From molecular diagrams to material properties: investigation of data-driven tools and strategies in the field of Chemoinformatics



Andre Paulo Pereira de Oliveira Frade

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

Acknowledgements

English version

After four amazing years, another stunning journey comes to an end. This thesis represents hard work, dedication, and an enormous passion for science and learning new things. Nevertheless, a beautiful story of enjoyment and self-growth can also be read in between the lines. By no means do I claim to have accomplished it on my own. In fact, this adventure was built and shared with many other characters who played their very own key role into making me the person I am today and supporting me towards this great achievement. Thus, no one should really read this thesis without realising all of its direct or indirect contributors. I would like to express my gratitude to all.

To my academic supervisor Professor Richard Cooper who was happy to go out of his way so I could pursue my passion and ideas. A truly supportive supervisor that has always embraced my thoughts and helped me to develop them into good science. The last three years were a fantastic and intense learning curve, and I do appreciate all of his amazing and invaluable daily guidance, every step of the way.

To my industrial supervisor Dr. Patrick McCabe for his amazing support and thoroughness, who helped me develop my sense of critical thinking and scientific rigour.

To my grandmother Ligia, who I primarily dedicate this thesis to. In her eyes, this ''faff'' might even be a bit redundant, because to her I have always been a little doctor of all sorts – and it is general knowledge that grannies know best!

To my beloved parents that have always encourage and supported me in all my endeavours. The two people in this world that always went above and beyond to give me all the opportunities I have always wished for… no questions asked. For this I will always be grateful.

To my adored sister that has always stood by me. To my caring aunts, uncles, and cousins. To this entire bunch that has always believed in me, much more than I have ever believed in myself - a truly united and supporting family where my success is their success.

To my John for his daily patience and support. For believing that I can do anything I put my mind to and encouraging me every step of the way. With this one alone I learnt enough to earn at least another couple of diplomas in the school of life.

To my friends that made this roller-coaster of emotions more enjoyable.

To wine, for a smoother and happier journey.

And lastly, to myself, for giving it a chance, rolling up the sleeves, and put all the hard work and brainpower into this. I could not be prouder of myself.

Versão Portuguesa

Após quatro anos incríveis, mais uma fantástica jornada termina. Esta tese representa trabalho árduo, dedicação, e uma enorme paixão por ciência e aprendizagem. No entanto, uma bonita história de satisfação e crescimento pessoal também pode ser lida nas entrelinhas. De forma alguma alego ter realizado tudo isto sozinho. Na verdade, esta aventura foi construída e compartilhada com muitos outros personagens que desempenharam um papel fundamental no desenvolvimento da pessoa em que me tornei e no apoio desta grande conquista. Assim sendo, ninguém deveria ler esta tese sem primeiro se aperceber de todos os seus diretos ou indiretos contribuintes. Gostaria de expressar a minha gratidão a todos.

Ao meu supervisor académico, Professor Richard Cooper, que saiu da sua zona de conforto para que eu pudesse seguir a minha paixão e implementar as minhas ideias. Um supervisor verdadeiramente apoiante, que sempre recebeu as minhas ideias de mente aberta e que me ajudou a torná-las em ciência de qualidade. Os últimos três anos resumem-se a uma curva de aprendizagem fantástica e intensa. Por tudo isto, lhe agradeço a sua orientação incrível e inestimável a cada etapa do percurso.

Ao meu supervisor industrial, Dr. Patrick McCabe, pelo seu incrível apoio e meticulosidade, o que contribuiu para a expansão do meu pensamento crítico e rigor científico.

À minha avó Lígia, a quem dedico esta tese. Aos seus olhos, tudo isto chega a ser um pouco redundante porque acredito que sempre me tenha visto e gabado como sendo um pequeno doutor, e é de conhecimento geral que as avós é que sabem.

Aos meus queridos pais que sempre me incentivaram e apoiaram a alcançar todos os meus sonhos. As duas pessoas neste mundo que sempre fizeram de tudo para me

proporcionar todas as oportunidades que sempre desejei. Por isto ser-lhes-ei sempre e imensamente grato.

À minha irmã que sempre esteve do meu lado e me apoiou. Às minhas tias, tios e primos. A uma família que sempre acreditou em mim, mais até do que eu acreditei em mim mesmo. Uma família verdadeiramente unida e altruísta, onde o sucesso é celebrado e partilhado.

Ao meu John, pela sua paciência e apoio diário. Por acreditar que eu posso alcançar todos os meus sonhos e encorajar-me a cada passo do caminho. Só com este aprendi o suficiente para ganhar pelo menos mais dois diplomas na escola da vida.

Aos meus amigos que tornaram esta montanha-russa mais divertida.

Ao vinho, por uma viagem mais tolerável e feliz.

E, por último, a mim mesmo, por agarrar esta oportunidade, arregaçar as mangas e trabalhar arduamente neste projeto. Não poderia estar mais orgulhoso de mim mesmo.

Table of Contents

Acknowledgements 3

Table of Contents 7

Abstract 12

Abbreviations 14

List of figures 16

List of tables 18

Chapter 1 - Introduction 19

1.1 Brief overview of QSAR/QSPR studies 19

1.2 The challenges of the discipline 21

1.3 The aims of this project 25

1.4 The crystallizability dataset 28

1.5 The structure of this thesis 29

Chapter 2 – Understanding the problem 32

2.1 Introduction 32

2.1.1 Problem formulation 33

2.1.2 Target labels 33

2.1.3 Molecular representations 34

2.1.4 Machine learning approaches 36

2.1.5 Wrap up 38

2.2 Results 38

2.2.1 Problem formulation 38
2.2.2 The datasets 41
2.2.3 The models 45
2.3 Conclusion 49
2.4 Methods 51
Software and Databases 51
Data Collection 52
Labelling Algorithms 52
Labelling Studies 54
Data Processing 55
Molecular Representation Studies 56
Dataset Visualisation 56
Machine Learning Models 57
Chapter 3 – Improving chemical datasets 58
3.1 Introduction 58
3.1.1 Molecular descriptor design 58
3.1.2 Intermolecular interaction propensity 60
3.1.3 The assessment of feature usefulness 61
3.1.4 Wrap up 62
3.2 Results 63
3.2.1 Feature library implementation 63
3.2.2 Feature usefulness assessment 65
3.3 Conclusion 68
3.4 Methods 70
Feature Library Implementation 70

Feature Redundancy Studies 70
Feature Relevance Studies 71
Chapter 4 – Increasing the practical value of models 73
4.1 Introduction 73
4.1.1 The vastness of chemical space 73
4.1.2 The stationarity assumption of statistical learning 74
4.1.3 The applicability domain of data driven models 76
4.1.4 The outlier detection problem 79
4.1.5 Wrap up 81
4.2 Results 82
4.2.1 Chemical diversity studies 82
4.2.2 Generalisability Tool 85
4.2.2.1 Generalisability tool description 85
4.2.2.2 Generalisability tool implementation 85
4.2.3 Confidence Tool 92
4.2.3.1 Confidence tool description 85
4.2.3.2 Confidence tool implementation 85
4.2.4 Applicability domain 99
4.3 Conclusion 102
4.4 Methods 104
Crystallizability Dataset & Model 104
Chemical Diversity Studies 106
Chapter 5 – Deriving chemical insights 107
5.1 Introduction 107
5.1.1 The curse of dimensionality 107

5.1.2 The feature selection process 108
5.1.3 Interpreting QSPR models 110
5.1.4 Wrap up 112
5.2 Results 113
5.2.1 Feature Selection 113
5.2.1.1 Feature selection pipeline description 113
5.2.1.2 Feature selection pipelin implementation 120
5.2.2 Model Interpretability 124
5.2.2.1 Model interpretability pipeline description 124
5.2.2.2 Model interpretability pipeline implementation 130
5.3 Conclusion 139
5.4 Methods 140
Genetic Algorithm 140
Dimensionality Reduction Performance 142
Knowledge Extraction Validation 142
Chapter 6 - Conclusion 144
Chapter 7 - Future Work 147
Chapter 8 - Publications 149
References 150
Supplementary Information 165
A. Molecular representations 165
B. Estimator and hyperparameter optimisation results 167
C. Crystal habit labelling algorithm visualisation 169
D. Extreme outlier threshold selection 170

E. Functional groups, α, and β values currently implemented in the IIP library ... 173

F. Feature redundancy studies ... 174

G. Feature relevance studies ... 177

H. List of *in-built* multivariate outlier detection estimators ... 179

I. Generalisability studies ... 180

J. Confidence studies and tool applications ... 182

K. Redundancy studies ... 185

L. Evolution of the dataset's dimensionality and predictive power, as it flows through the feature selection pipeline ... 189

M. Top most relevant features driving the prediction problem ... 190

Abstract

The field of Chemoinformatics has enabled QSAR/QSPR predictive models useful for the rapid virtual assessment of compounds or mining knowledge that may contribute towards the understanding of chemical systems. However, scepticism remains about the practical value of these approaches: models often exhibit limited predictive power, low generalisation capacity, and vague interpretability. More broadly, the field faces a general lack of consistency of results and procedures. This thesis addresses some of these challenges and explores new solutions to the field. First, standard protocols are used to explore the challenges of building chemical datasets suitable for statistical learning and understanding limitations of the predictive power of models. Here, the challenges of dataset curation and improvements to molecular structure description are explored. A library of new descriptors is built and assessed, which cleanly captures some information about crystallizability, but did not lead to measurable performance improvements of the models studied. The lack of relevant information content of some chemical datasets is reviewed, and strategies developed to increase the value of low-accuracy models: I report the generalisability and confidence tools and methodologies designed to restrict model predictions to inputs within their applicability domain. Illustrative implementations led to performance increases of up to 11% points of our predictive models. Finally, attempts are made to derive insights about structure-property relationships from chemoinformatics models. I develop a feature selection strategy which identifies the descriptors that influence a model's predictions, and an interpretation method that identifies useful changes to a molecule to control predicted properties. Insights such as the correlation between the higher crystallization propensity of molecules and the lower volume and flexibility of molecules are identified. This work contributes towards the motivation and development of general-purpose open-source

solutions which are key to add consistency in the field, increasing trust, and leveraging QSPR models to their full potential.

Abbreviations

abod	Angle-Based Outlier Detection
AD	Applicability domain
ANMI	Normalised Mutual Information adjusted for chance
API	Application Programming Interface
BFDH	Bravais-Friedel-Donnay-Harker
CCDC	Cambridge Crystallographic Data Centre
CSD	Cambridge Structural Database
CT	Confidence tool
EPSRC	Engineering and Physical Sciences Research Council
FS	Feature selection
GA	Genetic algorithm
GT	Generalisability tool
HBND	Hydrogen bond network dimensionality
hbos	Histogram-based Outlier Score
if	Isolation Forest
IIP	Intermolecular interaction propensity
kNN	k Nearest Neighbours
lof	Local Outlier Factor236
MI	Mutual information
MOD	Multivariate outlier detection
MOE	Molecular Operating Environment
NA	Not Applicable

NB	Naive Bayes
NN	Neural network
ocsvm	One class support vector machine
OECD	Organization of Economic Co-operation and Development
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
RBF	Radial basis function
RF	Random forest
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry System
SVM	Support vector machine
t-SNE	t-Distributed Stochastic Neighbor Embedding
VSA	van der Waals surface area

List of figures

Figure 1 Hydrogen bonding possible interactions and resulting network dimensionality.... 39

Figure 2 Main crystal morphology representations.... 40

Figure 3 Polymorphic and monomorphic packing representation diagrams.... 40

Figure 4 Visualization of the different datasets.... 45

Figure 5 Confusion matrices.... 48

Figure 6 Learning curves.... 49

Figure 7 Representation of the electrostatic potential energy plotted on the van der Waals surface of N-methyl acetamide.... 77

Figure 9 The generalisability and confidence tools within the QSPR prediction pipeline..... 82

Figure 10 Chemical diversity study results for the crystallizability case study.... 83

Figure 11 The generalisability tool.... 85

Figure 12 The generalisability tool within the QSPR prediction pipeline.... 88

Figure 13 Deployment scenario results on the crystallizability dataset.... 91

Figure 14 The confidence tool.... 94

Figure 15 A schematic representation of the confidence tool within the QSPR prediction pipeline.... 95

Figure 16 Confidence study results for the HBND dataset.... 97

Figure 17 The feature selection pipeline.... 113

Figure 18 Feature redundancy minimisation algorithm.... 115

Figure 19 The genetic algorithm.... 119

Figure 20 Redundancy study results for the crystallizability dataset.... 121

Figure 21 The model interpretability pipeline.... 125

Figure 22 Schematic representation of class probability density function profiles of a hypothetical feature..... 129

Figure 23 Magnitude study results for the crystallizability case study. 132

Figure 24 Feature alternative study results for the crystallizability prediction problem.... 135

Figure 25 Feature alternative study results for the HBND prediction problem. 136

Figure 26 Molecular diagrams of the Molecule A (non crystallizable class) and Molecule B (crystallizable class) pair example. 137

Figure 27 Extreme outlier threshold study results..... 171

List of tables

Table 1 Problem formulation summary. 41

Table 2 Labelling study results. 42

Table 3 Molecular representation study results. 44

Table 4 Summary of final and optimised QSPR models. 46

Table 5 Intermolecular interaction propensity feature names and definitions. 64

Table 6 IIP feature redundancy results for the crystallizability case study. 66

Table 7 IIP feature relevance results for the crystallizability case study. 68

Table 8 Generalisability study results for the crystallizability case study. 90

Table 9 Results per round of the SVM model for HBND prediction, subject to a sequence of seven confidence thresholds. 98

Table 10 Generalisability & confidence tool settings adopted for different case studies .. 101

Table 11 Applicability domain studies. 101

Table 12 Details of the test data used under the controlled deployment scenario 106

Table 13 Evolution of the crystallizability dataset's dimensionality and predictive power, as data flows through the feature selection pipeline. 122

Table 14 Top most relevant features driving the crystallizability prediction problem. 123

Table 15 Data assurance results for the crystallizability case study. 131

Table 16 Directionality study results for the crystallizability case study. 133

Table 17 Validation of the direction of the influence of each descriptor applied to the Molecule A and Molecule B pair example. 138

Chapter 1 - Introduction

In *Chapter 1*, the field of Chemoinformatics is introduced. The size and potential of the discipline are demonstrated and some of its current challenges are also exposed. The main aims of this DPhil project are discussed and the structure of this thesis is outlined.

1.1 Brief overview of QSAR/QSPR studies

Arguably, one of the primary goals of Chemistry is to develop sufficiently powerful theoretical understanding to enable us to explain the physical microscopic and macroscopic properties of chemical entities, as well as to anticipate what will happen when those entities are exposed to or combined under different conditions. Accordingly, different and complementary approaches to Chemistry often come together to solve particular problems, such as the identification of molecules and materials with specific properties and behaviours. Three of the main chemistry disciplines that play a key role towards this aim are Quantum Chemistry, Force Fields and Chemoinformatics.[1]

Chemoinformatics is a multidisciplinary field that emerged from the intersection of chemistry, computer science and information technology to enable the possibility of performing virtual experiments extended to an infinitely vast chemical space.[1] Ever since its foundation, the discipline has shown great potential, delivered great impact, and expanded into several specialised branches.[1,2]

From early stages, researchers established the idea that chemical structures directly influence the properties and activities they exhibit.[3] A significant area of Chemoinformatics is dedicated to quantitative structure-property relationship (QSPR) and quantitative structure-activity relationship (QSAR) studies.[4] Unlike other computational chemistry approaches, QSAR/QSPR studies are not based on natural laws nor do they aim to model

physical reality.[1,4] Instead, they help researchers make sense of large volumes of experimental data, which becomes particularly useful when complex or unclear systems and relationships cannot be easily modelled by physical-based methods.[1] Instead, data driven approaches are used to identify and rationalise the relationships between the structural features and chemical or biological properties of compounds. Such relationships are formally delivered as predictive models, which are conventionally produced by a two-step process[5]. In the first stage, molecular structures and property values are converted into mathematical objects suitable for the structure-property mapping task. The second step consists of finding the function that best captures and generalises that relationship. The resulting QSPR/QSAR models act as powerful tools used for the rapid *in silico* assessment of novel molecules, screening of virtual chemical libraries and guiding the design of property/activity-optimised compounds.[4,6]

Chemoinformatic models have already enabled the accurate prediction of numerous molecular properties and interactions,[7] such as pKa values[8,9] or polarizability.[10,11] However, their success seems to also extend to the prediction of material properties from knowledge of molecules alone, with no explicit consideration of the organisation of the crystal structure that will form. Examples include melting point,[12,13] boiling point,[14–16] density,[17,18] viscosity,[19,20] solubility,[21–23] partition coefficients,[24,25] refractive indices,[26,27] dielectric constants,[28,29] surfactant properties,[30,31] crystallization propensity,[32,33] to mention a few.[14,34–36]

QSPR and QSAR models were primarily developed to support the development of drugs, helping with the identification of promising compounds in the early stages of the pipeline to increase the speed with which drugs reach the market.[37] So far, QSPR strategies have contributed to the identification and development of numerous drugs which are currently in clinical trials or have already reached the market.[38–40] However, chemoinformatics approaches and models have great potential and started to be successfully

adopted in the research and development pipelines of many scientific fields and industry sectors, including biology, agriculture, medicine, chemistry, and material sciences.[37,40,41–48,49] Thus, any new contribution to the field has potential for impact across a wide range of areas and disciplines.

1.2 The challenges of the discipline

Modern chemoinformatics heavily relies on data driven approaches, which have seen great progress in the last few decades empowered by the fast-growing amount of publicly available data, advances in modelling procedures and inference techniques, and an increase in the computational power of machines.[4,6,50] This transition has enabled computational tools in the domain of machine and deep learning, that are more powerful than ever.[4,6,50] Yet, many of the tools, techniques and knowledge it employs are active areas of research and despite its achievements, there are still many questions to address, challenges to overcome and procedures to rationalise.

Chemoinformatics predictive models are very appealing because they promise to deliver detailed information about existing and hypothetical compounds in a fraction of the time and resources required by traditional experimental methods or first principle electronic structure approaches.[51–56] Moreover, these models may also be used as an alternative approach to research, contributing towards the understanding of the chemical systems and advancement of chemistry as a discipline.[57,58] However, the disappointing results in certain situations discourage practitioners to fully embrace QSAR/QSPR studies, who argue that these have yet to demonstrate their robust and consistent ability to add value.[58] In fact, there has been a general rise of concern that the majority of published models do not meet the requirements for a reliable practical use.[59–63] The scepticism around the value of these approaches stems from a number of reasons, most of which are summarised below.

Individual models often exhibit limited performance during production, lack of generalisability to unseen examples after deployment, and/or vague interpretability for knowledge extraction. The predictive power of a model is primarily dependent on the success of the mapping task between independent features and target variable. This exercise alone relies on a number of factors from which the limited performance can originate. These include incorrect problem formulation or class definition, errors in data, incomplete representation of molecular structures, high dataset dimensionality, or suboptimal choice of base estimator and hyperparameters.[60,61,64,65] In fact, the complexity of the systems being modelled and the absence, or inaccuracy, of property-relevant descriptors and target labels may explain why many properties remain so difficult to predict.[61] On the other hand, the lack of generalisability to unseen examples often stems from the inadequacy of inputs that do not meet the requirements to be reliably processed by the model.[66] Finally, molecular features span across different levels of abstraction and the success of knowledge mining is limited by the level of interpretability of highly predictive features.[67,68] A review of the manifold reasons for errors in QSAR/QSPR studies can be found in *Dearden et al.*[63]

Owing to the lack of trust in individual models, important steps towards regaining the faith in the field have begun. Many best practices have been suggested to address most of the challenges mentioned above and provide general guidelines towards the production of robust, reliable and reproducible models.[63,69–72] Moreover, the OECD has also established a few principles to facilitate the consideration of QSPR models for regulatory purposes.[73] And finally, a big push towards moderating the expectations around any QSAR/QSPR results and conclusions has also clarified that instead of replacing the domain knowledge of scientist, these should be strictly viewed as complementary information to support the decision-making process in research pipelines.[62,74]

In short, the production of robust and reliable QSAR/QSPR models must consider the following aspects: a) Models are only as good as the data used to produce them.[61,75,76] Data curation improves the quality of the dataset for the subsequent analysis through the standardisation of raw data and removal of any inconsistency, error, or duplicated information along the process. b) Model validation aims to assess the predictive power of the dataset and helps to ensure that the model will be able to generalise to unseen examples.[69] This can be done externally with a test set, and internally using a separate validation set (if data are abundant) or the training set under a framework that allows statistically viable results (if access to data is limited). For example, the popular *k-fold cross-validation* method splits the training set into k independent groups of similar size. Then, each group is used to assess the performance of the model built from the remaining folds, and the average of model performances becomes a robust performance metric.[4,69,77] Model validations is a crucial step to diagnose overfitted models. Overfitting is a concern of particular importance due to the frequency and easy with which it can occur.[78] That happens when models learn the characteristics of the training data too well, including their noise or random fluctuations, making them unable to generalise to unseen validation or test examples. Overfitted models often have no practical use. c) Furthermore, the machine learning task can be more objectively and efficiently described through the removal of useless features and harmful examples.[79,80] Feature selection identifies the set of most useful features that drive the prediction task, redrawing the focus of computational resources to the relevant subset of data. d) Outlier detection methods reveal data points that differ significantly from the majority of the observations, which may diminish the quality of models and bias the results. e) Also note that QSAR models are not universally applicable and predictions may fail under certain conditions.[79,81] The scope or applicability domain of a model refers to the collection of unseen inputs that it can process reliably and predict with

confidence. The identification of suitable inputs is an important step towards the trustworthy implementation of models, but it is often outside the control of the developers of a model once it goes into wider use or 'production'. f) Finally, models should be interpretable.[57,58] This step relies on the creation and adoption of features with meaningful definitions. The extracted insights not only play a critical role towards increasing the transparency and trust of QSPR studies, but could help practitioners to propose new mechanistic hypotheses about the cause-effect of structure-property relationships.

In a wider context, the field faces a general lack of consistency of results and procedures.[61] The variability of the model construction schemes may give rise to different outcomes and conclusions.[61,82] For example, model training is an optimisation problem guided by the minimisation of the average model error or maximisation of its performance. However, evolving models based on single metrics has already shown the possibility of rendering several solutions to the same prediction task with very similar performances that nevertheless seem to act quite differently on the same unseen input.[83] Lastly, the vast majority of machine learning algorithms are well understood, parameterised, and their implementations are available in open source software. This is untrue for chemoinformatics tools and methods. Despite the extensive description of QSAR/QSPR solutions across the literature, only very few implementations are easily accessible, well documented and maintained. Furthermore, reported solutions often have a narrow scope or are context dependent. In turn, practitioners must rely on their literature interpretation and chemoinformatics knowledge to recreate their own versions of each method.[57,79,81] Unfortunately, this is easily followed by poor statistical tool choices and/or oversimplification of important assumptions, which may compromise results and generate further controversy. To this end, many efforts have been taken to add structure and standards to the field, but

the wide variety of methods and their suitability under different contexts still generates confusion amongst researchers and inconsistency of results across the literature.[57,79,81]

1.3 The aims of this project

Chemoinformatics approaches become more accurate and widely accepted as their anchor fields evolve. The potential, size, and impact demonstrated thus far has motivated the need to add structure to the discipline through the clarification of concepts, establishment of standards, and provision of general-purpose open-source implemented solutions. In collaboration with industrial partners, this DPhil research project aims to address some of the above Chemoinformatics predictive modelling challenges and suggest new alternative and general solutions the field. We conduct our experiments through the study of a series of solid-state material properties, namely hydrogen bond network dimensionality, crystal habit and polymorphism. These were selected due to their high relevance in drug development pipelines, the lack of methods for practical *in silico* determination, and the access of relevant data. Currently, these properties still entail resource and time-consuming experimental studies, as to our knowledge no QSPR predictive solutions have yet been reported. In addition, the models span a range of predictive accuracies, providing a wide-ranging and realistic set of example problems to study.

Hydrogen bond network dimensionality (HBND) describes how hydrogen-bond intermolecular interactions extend in a three-dimensional structure. The expansion of the network is guided by the set of available hydrogen bonding groups in a molecule and their allowed interactions.[84] The resulting 'dimensionality' is thought to be a major cause of anisotropic interactions in crystal structures due to its directional nature.[85,86] Although their impact is not well characterised, dimensionalities often act as valuable complementary

information to the study of properties that are directly influenced by slip plane arrangements in crystals, such as crystal stability, mechanical behaviour and tabletability performance.[84,87,88] *Bryant et al.*[84] recently described an automated method to assign the dimensionality of a hydrogen bond network from solved crystal structures, and have demonstrated its effectiveness on multiple drug systems by comparison with tabletability data. For example, crystal forms with a three-dimensional hydrogen bond network are expected to exhibit an elastic deformation upon mechanical stress, *i.e.* particles return to their original position after being compressed.[89] From a drug formulation perspective, such crystals would produce porous and unstable tablets, therefore they would require to be mixed with compressible excipients. This is the case of acetaminophen, phenacetin, piroxicam and others.[90] Unbidden surprises like these entail further drug formulation challenges and delays, so predicting HBND in early stages would speed up the drug development process. However, the reliance on solved crystal structures as input limits the large-scale implementation of the tool. Crystals of compounds of interest are rarely available, obtaining them is resource and time consuming, and crystal structure predictions are computationally expensive and still not reliable enough for such application.[60,91] Currently, the HBND of a compound can only be determined analytically and crystalline samples are required.

Crystal habit refers to the physical shape of a molecular crystal. Habit is determined by the directions in the crystal with the slowest growth rates.[92] Crystal morphology not only affects various stages of the manufacturing processes of molecular materials, such as filtration, milling or compaction, but also the physicochemical properties and/or bioavailability of the solid product.[93] The desired crystal morphology depends on its application. For example, needle shaped crystals are often undesirable in pharmaceutical

settings for making molecular materials hard to process and formulate, not least because they clog filters and do not 'pour' easily.[92] Under these circumstances, crystal engineering measures are adopted to encourage the crystals of compounds like paracetamol and other active ingredients to growth into alternative shapes.[93] From the drug development perspective, crystal habit prediction helps to anticipate potential hurdles in later phases of the drug development pipeline. Crystal morphology studies heavily rely on experimental determination or crystal structure prediction combined with complex and computationally expensive mechanistic models of crystal growth.[94] The pioneering theory for the prediction of crystal morphology was formulated by Bravais, Friedel, Donnay, and Harker (BFDH) as the first of its kind to link the macroscopic structure of crystals to their internal molecular arrangements.[95,96] The theory is based on the assumption that the growth rate of a crystal face is inversely proportional to its interplanar distance. Consequently, faces with largest interplanar distances grow the slowest, becoming the largest and morphologically most relevant. However, the BFDH theory does not take into account the statistical thermodynamics of the crystal interfaces or any external factors.[97] Due to its entire reliance on geometry, the theory seems to only be suitable for the habit prediction of simple isotropic crystals.[97] Despite these limitations, this is still considered one of the most reliable approaches by sectors like the pharmaceutical industry.[92,97]

Crystal polymorphism refers to a compound's tendency to crystalize in different structural forms. Each packing alternative has its own physicochemical, pharmacodynamics and pharmacokinetic set of properties.[98] Drugs like paracetamol, cortisone and ritonavir are examples of compounds for which only one of the polymorphic forms is pharmacologically viable, either due to formulation, stability or solubility reasons. These and others are reviewed in *Censi et al.*[98] From the drug development perspective, it is critical to identify

whether a compound exhibits polymorphism and rationalize protocols and conditions to ensure formulation control.[99] Therefore, a model able to predict whether a compound exhibits polymorphism would be useful in drug development. Predicting the polymorphism propensity of a compound has been previously attempted. Most of the models rely on crystal structure predictions, where the relative stability of predicted polymorphs is based on lattice energy and entropy calculations.[100,101] Unfortunately, polymorph energetic differences are usually very subtle and the output prediction space is generally too large to allow useful conclusions. Therefore, the pharmaceutical industry still relies on experimental solid form screening of compounds, whose associated techniques are time consuming and non-exhaustive.[99] Therefore, it is critical to identify whether a compound exhibits polymorphism, and rationalize protocols and conditions to ensure formulation control.

1.4 The crystallizability dataset

In this project, most experiments and methods were conducted or developed using a large and well-behaved dataset. Wicker *et al.*[32] successfully modelled the crystallization propensity of organic compounds, using a cleverly designed dataset where molecules were assigned to one of two classes - crystalline or non-crystalline – based on their ability to form good quality crystals that would be suited for synchrotron X-ray diffraction experiments. In short, two databases were used for the retrieval of molecules. The ZINC database[102] contains biologically relevant and purchasable compounds from many chemical suppliers. Moreover, the Cambridge Structural Database (CSD)[103] is almost entirely composed of crystal structures determined by single crystal diffraction, providing a reliable source of crystalline compounds. Following the original strategy, molecules were initially retrieved from ZINC. All of those that also occur in the CSD were assigned to the crystallizable class of compounds, with the exception of structures solely determined by powder diffraction. The remaining were

assumed to not form crystals and used as the non-crystallizable set. Molecules that do not occur in the CSD are not necessarily non-crystalline, but the bias introduced by this assumption was shown to be small. The retrieved data was further standardised, and filtered for drug like molecules. An extended version of the crystallizability dataset was obtained from our CCDC collaborators. The updated dataset contains 440,653 labelled examples, equally distributed between both classes. Molecules and labels were retrieved using the original strategy under the same assumptions, except that no drug-likeness filters were applied. This dataset was widely used throughout the project, as its size, diversity, and the high predictive power it enables make it well suited for the proof of concept, development, and validation of new methods and pipelines in the chemoinformatics context. The dataset was processed and the modelling framework (base estimator and hyperparameters) was optimised following the same approach as described in ***Chapter 2 / Methods / Data Processing*** and ***Chapter 2 / Methods / Machine Learning Models***.

1.5 The structure of this thesis

Each chapter of this thesis contains a short introduction that covers the relevant literature, concepts and tools to the problems and questions it addresses.

In *Chapter 1*, the field of Chemoinformatics is introduced. The size and potential of the discipline are demonstrated and some of its current challenges are also exposed. The main aims of this DPhil project are discussed and the structure of this thesis is outlined.

In *Chapter 2*, we investigate whether it is possible to predict material properties from information derived from molecular diagrams alone. The concepts of problem formulation, molecular representations, and machine learning estimators are further explored. In this section we follow standard model production protocols to explore the challenges of building chemical datasets suitable for statistical learning and understand the predictive power

limitations of QSPR models. The key outcomes include three curated datasets and three state of the art predictive models.

In *Chapter 3*, we address the challenges faced during chemical dataset construction and test whether there is room for the improvement of molecular structure description. We explore concepts such as molecular feature design and feature usefulness assessment. The key outcomes include a library of novel structural low dimensional features aimed to describe the intermolecular interaction propensity (IIP) of molecular structures.

In *Chapter 4*, we review the lack of relevant information content of some chemical datasets and design strategies to increase the practical value of limited performing models. Here, we cover the concepts of stationarity assumption, model generalisability, prediction confidence and applicability domain in detail. The key outcomes include the generalisability tool designed to restrict models to regions of chemical space where they hold their predictive power, and the confidence tool designed to identify confident predictions and increase the output quality of classifiers. This confidence tool is published in *Frade et al.*[104]

In *Chapter 5*, we attempt to derive chemical insights about structure-property relationships from chemoinformatics models. Here, we cover concepts such as feature selection and model interpretability. The key outcomes include a feature selection strategy that identifies the structural aspects that most drive the prediction of a particular property, and a model interpretation methodology that reveals how changes in the molecular structure may affect the predicted property value.

Finally, *Chapters 6, 7* and *8* correspond to the conclusion, future work and publications, respectively.

This project falls within the EPSRC research areas of Artificial Intelligence Technologies, Statistics and Applied Probability, and Computational & Theoretical Chemistry.

Pfizer and the Cambridge Crystallographic Data Centre (CCDC) are collaborators and part funders of this DPhil research project.

Chapter 2 – Understanding the problem

In Chapter 2, we investigate whether it is possible to predict material properties from information derived from molecular diagrams alone. The concepts of problem formulation, molecular representations, and machine learning estimators are further explored. In this section we follow standard model production protocols to explore the challenges of building chemical datasets suitable for statistical learning and understand the predictive power limitations of QSPR models. Finally, we conclude that the incomplete representation of molecular structures may be a major factor restricting the predictive power of our models to higher performances. The key outcomes include three curated datasets and three state of the art predictive models.

2.1 Introduction

Chemoinformatics models promise to deliver detailed information about compounds in a fraction of the time and resources required by traditional methods that often involve compound synthesis and experimental property determination.[6]

After problem formulation, building a QSPR model generally requires three components: 1) a dataset of compounds labelled with the value of the property to be predicted, 2) a molecular representation that is suitable for the prediction task, and 3) a model base estimator that can learn the mapping between molecular representations and property values. Each of these steps are described in outline below in general terms.

There has been a general rise in concern about published QSPR models. Some would argue that the majority do not meet the requirements for a reliable practical use.[59,60] The core of poor performance is often attributed to limitations in the data, whether it relates to

the poor quality of the labels [60,61] or to the poor choice of the molecular representation adopted.[64]

2.1.1 Problem formulation

The QSPR modelling process starts with the problem formulation step. From a machine learning perspective, QSPR problems are best expressed as supervised learning tasks. Here, the focus is to learn the link between explanatory features and a target variable from a collection of labelled examples. Moreover, the nature of the target property influences the selection of model estimators and performance measures that can be adopted. This factor alone defines the type of supervised learning task at hand. Regression problems are established when the property can assume values in a continuous scale, whilst classification occurs if the outcome is restricted to a limited set of values. Classification problems can be further categorised as being balanced or imbalanced, according to whether the various classes are represented by a similar number of examples.[6]

In the classification case, the quality of class definitions can have a tremendous impact on the success of the modelling task. Thus, classes must be defined in a way that is convenient for the intended application of the model, whilst still reflecting a natural separation of data.

2.1.2 Target labels

After successfully translating the chemical problem into a machine learning task, we move on to data collection, where a set of chemical systems and their corresponding property values are gathered.

The selection of molecules inevitably narrows down the focus of the dataset to a particular region of chemical space and sets the context under which any consequent

application finds appropriate usage and validity.[81] For example, a model trained on small organic molecules should be restricted to the prediction of other small organic molecules.

In addition, the accuracy of the labels greatly influences the quality of the dataset and of all results derived from it. Experimental labels may be expensive to get, often inconsistent across different sources, and negative examples are hardly reported in the literature.[105] Furthermore, in cases where algorithmic labelling is possible, the labelling consistency across the dataset is generally improved, but the algorithm's underlying assumptions and approximations may be a source of inaccuracies and biases. During property modelling, label values define the hidden patterns to be found in data. Thus, allowing sufficient mislabelled examples in the dataset may corrupt the entire learning exercise, especially when they heavily contradict the patterns revealed from high quality instances.

2.1.3 Molecular representations

In Chemoinformatics, molecular representations aim to describe the structures of compounds as mathematical objects that can be processed by computational and statistical methods. Chemical entities can be represented in various ways and different formats may be better suited to different tasks. A recent review of chemical representations can be found in Lo *et al.* [6] Popular examples include molecular graphs commonly used in pharmacophore analysis,[106,107] and molecular fingerprints often employed in similarity based virtual screening.[108,109]

The success of a given QSPR modelling task heavily relies on the information content of the molecular representation adopted. All representations are limited to a certain descriptive power. Different representations tend to capture different aspects of molecular structures, so it becomes crucial to find a format that represents the molecule well and highlights the key structural aspects that are most relevant to the property being predicted.

Descriptor vector representations are the most popular and the best suited for property modelling due to their modularity and scalability.[110,111] These consist of arrays of numerical features, such as molecular weight or number of hydrogen bond donors, which together build the molecule's unique profile.[111] Different types of descriptors are available. A thorough review of molecular descriptors can be found in *Khan et al*.[111] Two and lower dimensional descriptors are those that can be rapidly derived from molecular diagrams at low computational cost. These capture aspects such as the count and properties of atoms, bonds, functional groups, and structural fragments, as well as topological indices.[110,112] Despite their deterministic unambiguous computation, lower dimensional descriptors are often limited in their information content, which often lessens their discriminating power for different molecular conformations or subtle changes in molecular structures.[110,111] Given that molecules can co-exist in multiple conformations, these descriptors may be insufficient to fully describe a given property,[113] especially those to which conformational flexibility is highly relevant.[114,115] On the other hand, higher dimensional descriptors are able to capture the three-dimensional conformation of molecules and their interaction with the environment.[111,116] These often describe molecular structures in terms of their shape, steric, and electronic character.[39] Despite their high and relevant information content, these descriptors rely on the atomic coordinates of compounds, whose prediction is computationally expensive and cannot be guaranteed to correspond to the relevant conformation,[6,60,91,117] which can considerably increase the runtime of the algorithm without adding any useful contribution,[113] or even decreasing model performance.[118] In fact, lower dimensional descriptors were shown to perform significantly better than those of higher dimensions at several points.[119–122] The deterministic character and potential information content of descriptors are key factors to consider during their selection, as it will have

implications on property description, but also on model performance, robustness and stability.[115,123]

2.1.4 Machine learning approaches

Supervised machine learning is the engine behind chemoinformatics model production. Once chemical data is appropriately encoded, statistical methods and training techniques are used to find optimal functions that map feature vectors to properties of interest.[4]

A brief review of machine learning estimators can be found in *Lo et al*. [6] or *Mitchell et al.*[4] These methods became very prevalent due to their easy implementation and computational efficiency, as well as great performance on unseen examples. Some of the most commonly used procedures include random forests (RF), k-nearest neighbours (kNN), naïve Bayes (NB), artificial neural networks (NN) and their deep learning derivatives. Support vector machines (SVMs) are a popular class of statistical methods that allow classification problems to be formulated as a convex optimization task that can be solved in polynomial time, and generalise well from reasonably small datasets.[124,125] At their core, SVMs attempt to find linear hyperplanes for which the margin between data points of different classes is maximal.[126] SVMs can be extended to the identification of nonlinear patterns, where data points are transformed and projected onto higher dimensional spaces, such that classes become linearly separable. These transformations aim to emphasize the distances between points of different classes, and are often defined by kernels or feature projection functions like the Radial Basis Function (RBF) or the Polynomial kernel.[124] In addition to the choice of kernel, the method has two additional main hyperparameters: C and gamma. The parameter C defines the trade-off between the misclassification of training examples and the simplicity of the decision surface, and larger C values correspond to less regularisation strength. Finally,

gamma controls the radius of influence of each support vector, and larger gamma values correspond to smaller radii.

The suitability of each technique is primarily determined by the hypothesis to be tested, the size and distribution of the dataset and the complexity of the relationships to be found in the data.[4,33] However, the lack of prior knowledge of how structural features relate to a given property prevents the straightforward selection of the optimal method to be adopted for the mapping task. In practice, the same dataset is used to train many estimators in parallel and the one that performs the best is often selected.[33]

Fitted models must then be assessed for their ability to generalise to new unseen examples.[70,77] This is often achieved by testing the model's performance on a held-out set of labelled data and comparing predicted and true values. Classifiers are often assessed using *confusion matrices* from which other summary measures like *accuracy*, *precision*, and *recall* are derived.[127] A confusion matrix is a table that summarises the number of correct and incorrect predictions, broken down by class. Accuracy corresponds to the total fraction of correct predictions, recall is the fraction of examples of a given class that were correctly predicted, precision refers to the fraction of correct predictions out of all the examples estimated to belong to a given class., and the F1 score is computed as the harmonic mean of precision and recall. For balanced problems, accuracy and F1 score values should be the same. However, note that the accuracy measure can be misleading and lead to incorrect conclusions in class imbalanced problems. In such circumstances, F1 score works as a better alternative measure because it takes into account how data are distributed across classes.[127,128] Another useful model diagnostic measure is the learning curve, which shows the evolution of the validation and training performances for increasing amounts of training data. Learning curves can be used to diagnose how well the model is fitting the training data and generalising to unseen examples, as well as to derive upper bound accuracy limits given

the type of data and model at hand. These plots often help to assess models for bias, variance, and overfitting, providing good hints of how the model could be improved.

2.1.5 Wrap up

In this chapter we investigate whether it is possible to predict material properties solely using information derived from molecular diagrams of their constituents. Note that only materials with a single molecular component are studied.

Here, we follow standard model production protocols to explore the challenges of building chemical datasets suitable for statistical learning and investigate some of the predictive power limitations of QSPR models.

Each material property was investigated as a separate case study, yet all studies followed a general methodology that was primarily implemented and validated with the crystallizability dataset.

2.2 Results

2.2.1 Problem formulation

Three material properties form the basis for this work: hydrogen bond network dimensionality (HBND), crystal habit, and polymorphism. In this project we exclusively focus on classification problems. Property prediction tasks were thoughtfully formulated and classes were carefully defined according to scientific literature findings, expert domain knowledge, practical relevance and usefulness.

Hydrogen bond networks occur when copies of the same molecule can establish intermolecular hydrogen contacts between each other, when they self-organise in a crystal structure. The available hydrogen donor and acceptor groups present on each molecule guide the hydrogen network expansion and determine its dimensionality.[84] Networks may

extend in four distinct ways, enabling enclosed motifs, chains, sheets or expansion of the network in all directions. The term *dimensionality* refers to the number of dimensions in which the network extends via hydrogen bonds.[84] A schematic representation can be found in **Figure 1**. Thus, the hydrogen bond network dimensionality prediction can be formulated as a four-class classification task.

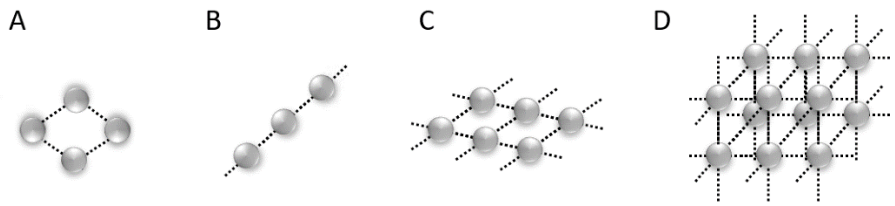


Figure 1 Hydrogen bonding possible interactions and resulting network dimensionality. A Zero-dimensional, 0D, comprised of enclosed motifs such as rings. B One-dimensional, 1D, chains. C Two-dimensional, 2D, sheets. D Three-dimensional, 3D, when the network expands in all possible directions.

The shape of a molecular crystal is mainly determined by its nucleation and growth process, during which the faces with the slowest growth rates become the largest and morphologically most important.[129,130] Despite the wide variety of shapes that crystals can adopt, the habit of single component crystals is usually conserved[130] and can be categorised into one of three industry relevant morphological classes: blocks, plates and needles. An example of each morphology type is schematically represented in **Figure 2**. Thus, crystal habit prediction can be formulated as a three-class classification task.

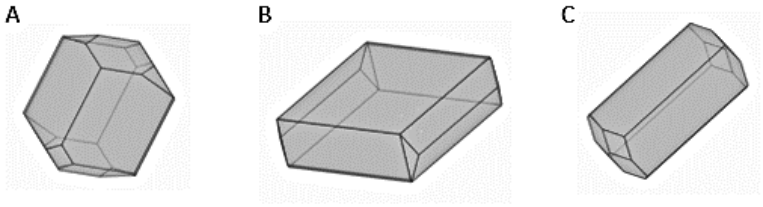


Figure 2 Main crystal morphology representations. A Block (CSD refcode AAPYPE). B Plate (CSD refcode AARBOX). C Needle (CSD refcode ABAFEQ). Crystal morphology was simulated using the CSD Mercury BFDH morphology method.

Some molecules are able to adopt more than one stable packing arrangement in their crystal structure, known as polymorphs. On the other hand, monomorphs are those that seem to always assemble themselves into the same low energy system. Whilst the polymorphic character of a molecules can be proved, the monomorphic character can only be assumed, and that is until a new packing form is discovered.[131] This imposes limitations on the accurate definition of the monomorphic class and on the quality assessment of the entire dataset. Yet, polymorphism prediction can be formulated as a two-class classification task. Examples of the behaviours expected from the molecules of each class are schematically represented in **Figure 3**.

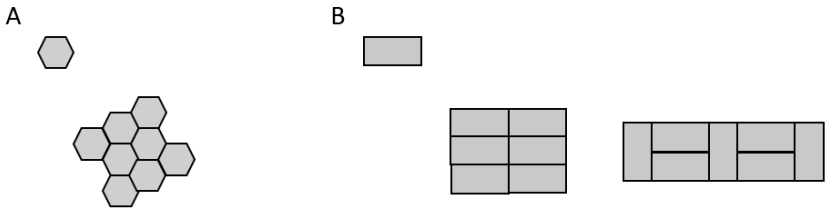


Figure 3 Polymorphic and monomorphic packing representation diagrams. A Monomorphic example, where only one packing form can be adopted. B Polymorphic example, where multiple packing forms can occur.

A summary of all problem formulations and respective classes can be found in **Table 1**. Accordingly, the scope of this project is narrowed to the investigation of QSPR classifiers. However, all the methods and analysis throughout this and the following chapters could be easily extended to other machine learning problem categories.

Table 1 Problem formulation summary.

	HBND	**Habit**	**Polymorphism**
Problem Formulation	4 class problem	3 class problem	2 class problem
Classes	0-dimensional 1-dimensional 2-dimensional 3-dimenaional	Block Plate Needle	Monomorphic Polymorphic

2.2.2 The datasets

Data were retrieved and labelled as described in ***Chapter 2 / Methods / Data Collection*** and ***Chapter 2 / Methods / Labelling Algorithms***. The labelling procedure was conducted using in-house developed and implemented algorithms that were customised to our problem formulations and class definitions, also promoting label consistency across dataset instances.

Labelling studies were carried out to assess the quality of each dataset. In this study, labelling error and instability were investigated. The labelling error rate estimate refers to the fraction of molecules of the dataset for which labels vary across resubmissions of the same crystal structure. The label instability rate estimate refers to the fraction of molecules of the dataset that could adopt more than one label value, *i.e.* those that possess a multi-label character. Note that the labelling error and label instability estimation strategies do not

apply to the polymorphism dataset due the definitions of its classes. In this case, the labelling error rate was replaced by the monomorphic assumption error estimation, and label instability as we define it does not exist in the polymorphism dataset. Labelling study results can be seen in **Table 2**.

Labelling errors may arise from the poor quality of crystal data and/or limitations and sensitivity of our label calculation algorithms. The study revealed that mislabelled events occur at a very low rate across all datasets. Furthermore, label variability within the same family of entries may occur due to labelling errors, but also in cases where the compound is polymorphic and its different crystal arrangements are reported to exhibit different label values.[84] The label instability rate is relatively low in the habit dataset but significantly higher across HBND entries. Finally, all visible errors and instabilities were corrected by discarding any molecule with clear label variability across its different entries.

Table 2 Labelling study results. The labelling error and instability rate estimates were calculated as described in ***Chapter 2 / Methods / Labelling Studies***. The labelling error rate value attributed to the polymorphism dataset corresponds to the monomorphic assumption error.

	HBND	**Habit**	**Polymorphism**
Labelling error rate	0.025	0.007	0.012
Label instability rate	0.117	0.055	0

Molecular representation studies were carried out to identify the molecular description that is best suited and most relevant to each property prediction task. In this study, only low dimensional molecular representations were considered. Despite any potential description limitation drawbacks, simpler structural descriptors tend to be deterministic in nature and can be rapidly derived from molecular diagrams at low

computational cost, ensuring the reproducibility and scalability of the resulting predictive models.[111] Three different molecular descriptions were implemented, each aimed to capture and focus on different structural aspects of compounds. The low-level representation was designed to highlight the atomic and sub molecular environments within each molecule. The mid-level representation aimed to capture the size, chemical composition, functionality and pharmaceutical relevant properties of each molecule. Finally, the high-level representation covers a wide and predefined range of structural aspects and physicochemical properties. Molecular representation details can be found in ***Supplementary Information A*** and results can be seen in **Table 3**.

This study revealed that whilst some molecular representations worked better than others, no available description contained all relevant information to enable a highly accurate mapping between the structural features and the property values of compounds. Taking the HBND case study example below, the mapping between structural high-level representations and network dimensionality values achieves an additional 7% in performance when compared to its low-level description counterpart, yet the best of the mapping scenarios could only achieve a maximum accuracy and F1 score of 54%. In summary, these findings confirm the limitations of predicting three-dimensional properties from low dimensional molecular representations. The results also illustrate the impact that the molecular description choice can have on the performance of a QSPR modelling task.

Table 3 Molecular representation study results. The suitability of molecular representations (high, mid, and low-level) under each case study context (HBND, crystal habit and polymorphism) was measured as the mean accuracy and F1 score obtained across three grid search cross validation routines, each using one of the model estimators considered (***Chapter 2 / Methods / Molecular Representation Studies***). Note that data was processed beforehand as described in ***Chapter 2 / Methods / Data Processing***.

Property	Mol. Representation	Overall Accuracy	Overall F1 score
HBND	**High-level**	**0.54**	**0.54**
	Mid-level	0.51	0.50
	Low-level	0.47	0.47
Habit	**High-level**	**0.48**	**0.48**
	Mid-level	0.47	0.47
	Low-level	0.47	0.47
Polymorphism	**High-level**	**0.61**	**0.61**
	Mid-level	0.60	0.60
	Low-level	0.58	0.58

The final HBND dataset resulted into 55389 dimensionality labelled examples - 14004 with 0D networks, 31314 with 1D networks, 7044 with 2D networks, and 3027 with 3D networks. The final crystal habit dataset resulted into 160161 habit labelled examples - 141376 blocks, 12629 plates and 6156 needles. The final polymorphism dataset resulted into 3133 labelled examples – 1194 monomorphic and 1939 of polymorphic nature. All instances were described by the high-level molecular representation, composed of 113 numerical descriptors.

Upon label and description calculation, datasets were appropriately processed, and subjected to t-Distributed Stochastic Neighbour Embedding (t-SNE), a nonlinear dimensionality reduction technique that embeds high dimensional datasets into two dimensions for visualisation purposes.[132] **Figure 4** displays the resulting two-dimensional

projections of each transformed dataset, which consistently show that data points are evenly distributed, visually suggesting that none of the datasets provide a simple separation between classes.

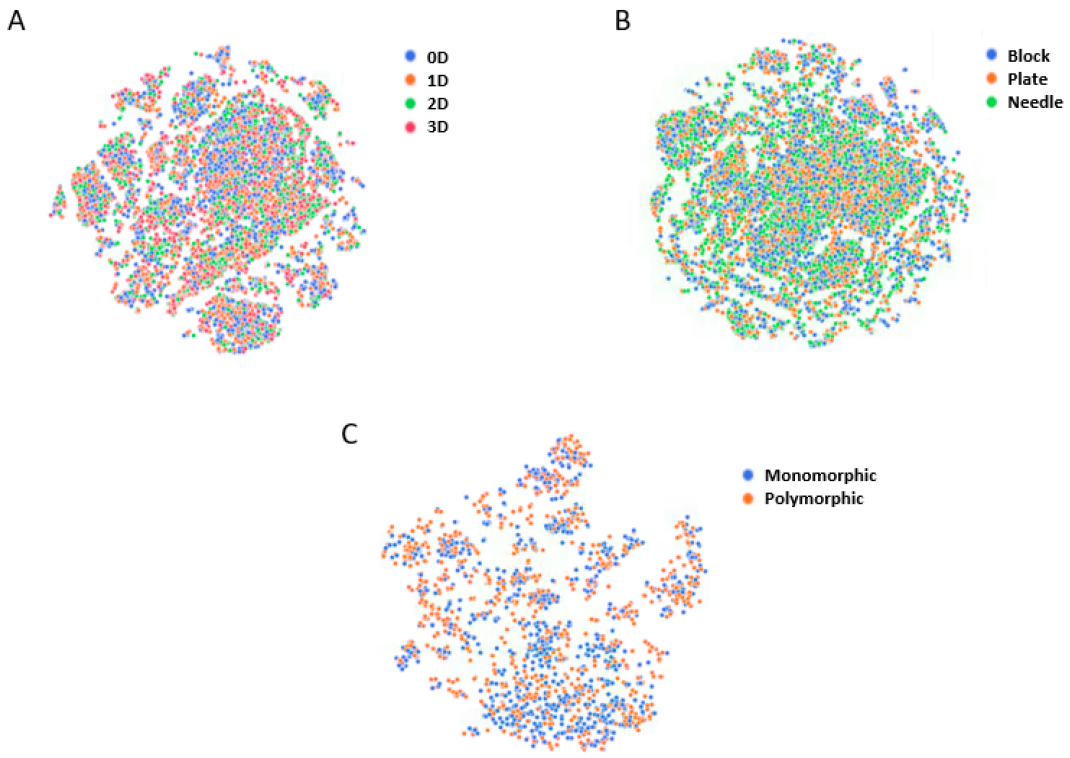


Figure 4 Visualization of the different datasets. A HBND dataset. B Crystal Habit dataset. C Polymorphism dataset. Each data point is an entry of the balanced dataset that has been compressed from its original dimensional space to two dimensions using the t-SNE method. Points are coloured according to the class they belong to. See ***Chapter 2 / Methods / Data Visualisation*** for more details.

2.2.3 The models

Processed datasets were used for model selection. Optimal models were evaluated on true and random test data. The prediction capability of each model and its ability to generalise to unseen examples were assessed by comparing true test performances and baseline reference values. Note that training and test sets were class balanced and

performances were measured using accuracy and F1 scores. The results can be found in **Table 4**.

All modelling frameworks achieved similar results, with a consistent slight overperformance of the SVM implementation across all three classification tasks (***Supplementary Information B***). Furthermore, all final models performed considerably better than random but still experienced some performance limitations, remaining far from achieving desirable accuracies. In summary, the predictive power of all three classifiers provides strong evidence that each dataset contains limited but relevant structural information for the modelling and prediction of each property.

Table 4 Summary of final and optimised QSPR models. The table summarises the optimal estimator, as well as baseline and test performances (on a class balanced test set) for each case study. Baseline performance refers to the performance that the model would reach if it was to perform no better than random. Models were obtained from high-level molecular representation datasets, processed as described in ***Chapter 2 / Methods / Data Processing***. Model selection (optimal estimator and hyperparameters) and evaluation was conducted as described in ***Chapter 2 / Methods / Machine Learning Models***.

	HBND	**Habit**	**Polymorphism**
Estimator	SVM	SVM	SVM
Baseline performance	0.25	0.33	0.50
Test Accuracy / F1 score	0.55 / 0.55	0.49 / 0.49	0.62 / 0.62

The best performing HBND predictive model was a multi-class implementation of the SVM estimator with a linear kernel (C=1, gamma=0.01), which attained an overall test accuracy of 55 +/- 1%, with accuracies per class varying between 47% and 66%. The corresponding confusion matrix can be found in **Figure 5A**. The classifier is able to detect each class with a performance considerably higher than random (baseline random accuracy value of 25% for class balanced test sets). We also notice that unlike other properties, HBND

misclassified examples tend to be assigned to adjacent classes, suggesting that the definition of network dimensionalities is a continuum with no well-defined boundaries between adjacent classes.

The best performing crystal habit model was also a multi-class implementation of the SVM estimator with an RBF kernel (C=10, gamma=0.001), which attained an overall test accuracy of 49 +/- 2%, with accuracies per class varying between 48% and 50%. The corresponding confusion matrix can be found in **Figure 5B**. The classifier is able to detect each class with a performance considerably higher that random (random accuracy value of 33% for class balanced test sets).

Finally, the best performing polymorphism predictive model was a binary implementation of the SVM estimator with an RBF kernel (C=1, gamma=0.001), which attained an overall test accuracy of 62 +/- 1%, with accuracies per class varying between 61% and 63%. The corresponding confusion matrix can be found in **Figure 5C**. The classifier is able to detect each class with a performance considerably higher that random (random accuracy value of 50% for class balanced test sets).

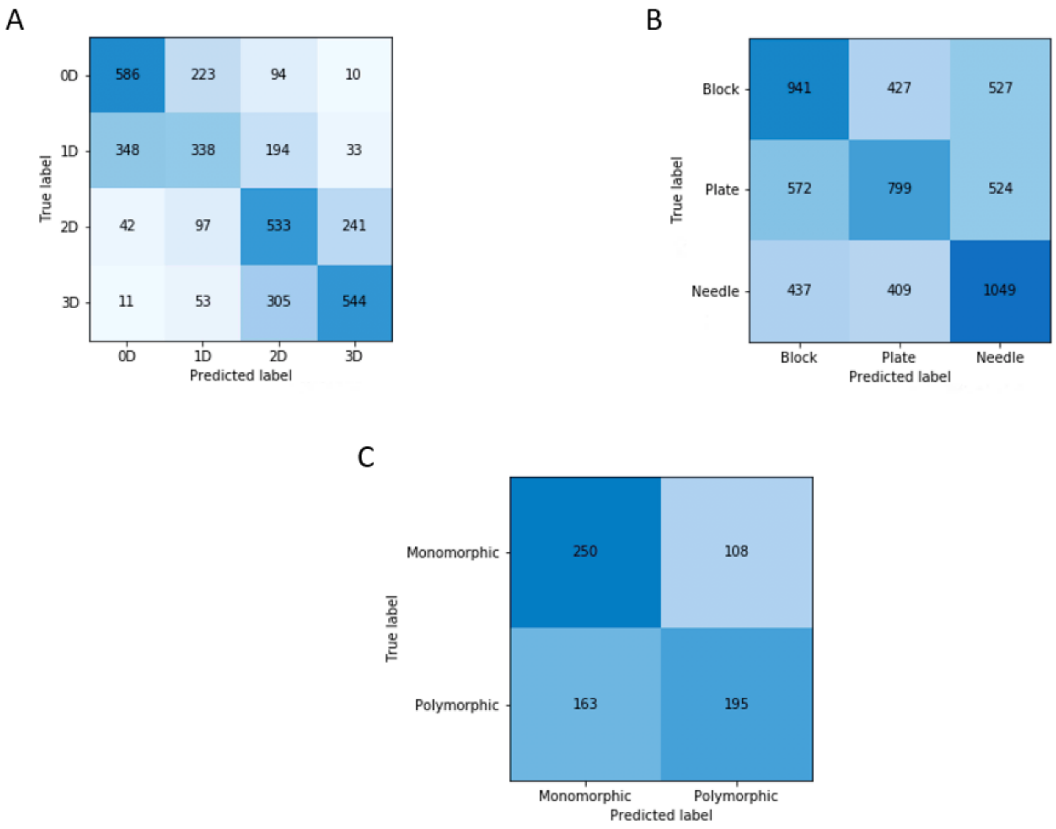


Figure 5 Confusion matrices. A Hydrogen Bond Network Dimensionality. B Crystal Habit. C Polymorphism.

Learning curve results show that models tend to generalise well to unseen examples, suggesting that no overfitting has occurred during the training stage. The consistent lack of convergence between training and cross validation score lines shows mid to high variance, whereas the relatively low accuracy level at which score lines plateau show mid to high bias, both suggesting that model performances could be improved. In short, a model with high variance and bias often represents a scenario where a high dimensional dataset is considered, but only a small subset of features effectively contributes to the mapping task. Furthermore, that subset is far from providing a full description of all structural relevant

aspects required to the accurate modelling of the target variable, which explains the limited performance of models.

Learning curves built from accuracy scores also provide upper bounds for how good a model can get using the set of descriptors considered. The upper bound corresponds to the accuracy at which both line scores would theoretically converge, which we estimate to be between 55% and 58% for HBND (**Figure 6A**), 50% and 57% for habit (**Figure 6B**), and 63% and 67% for polymorphism (**Figure 6C**), from each corresponding learning curve

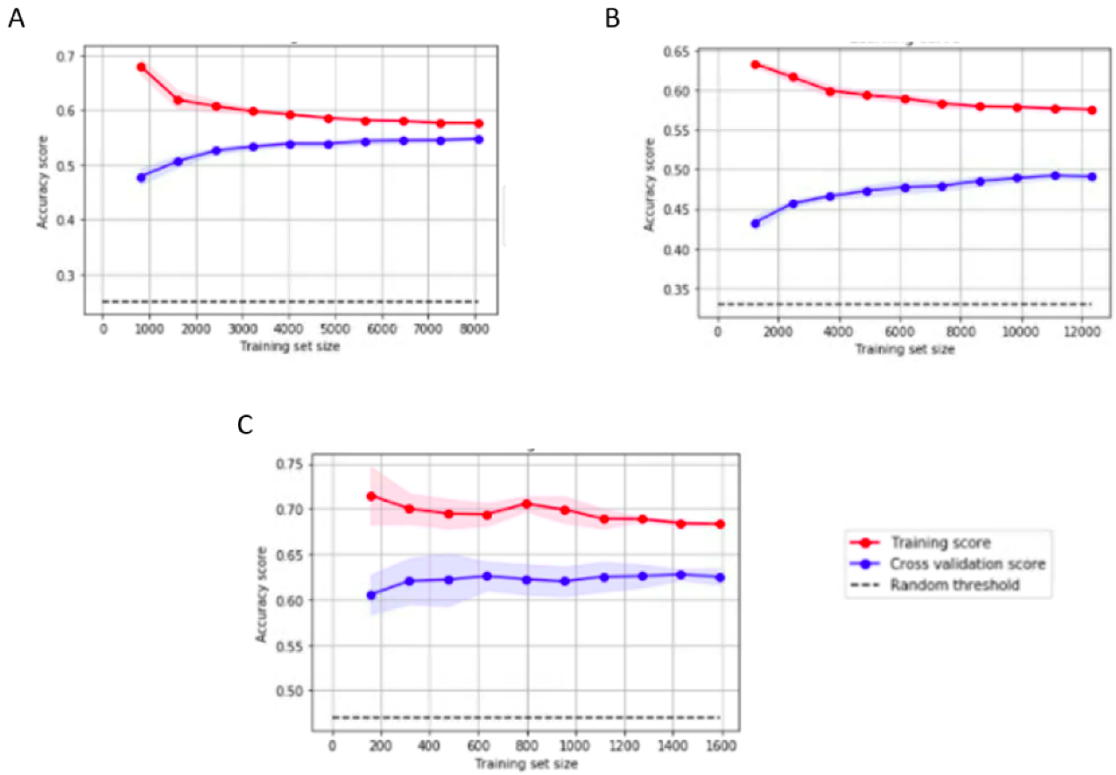


Figure 6 Learning curves. A Hydrogen Bond Network Dimensionality. B Crystal Habit. C Polymorphism.

2.3 Conclusion

Reliably predicting material properties using low dimensional structural information is a powerful strategy for the rapid and cheap assessment of existing and hypothetical

chemical compounds. Models like this are of interest to the pharmaceutical industry as they enable informed decisions on the most promising drug candidates in the early stages of drug development.

Here, we build, study and report three datasets and corresponding classifiers that confirm the possibility of estimating different material properties using low dimensional information derived from their molecular constituents. All models performed considerably better than random, yet their predictive power was limited. Note that limited performing models still find good use for the initial screening of large datasets. Furthermore, these provide us realistic systems that can be used to validate some of the solutions proposed in the subsequent chapters.

The production of QSPR models with significant predictive power requires high quality chemical data translated into a format that represents the molecule well and that emphasises the set of structural features that are most relevant to the property being modelled. The different factors that may account for such limitations have been discussed. Class definitions were mainly inspired by relevant literature (***Chapter 2 / Methods / Labelling Algorithms***), and model estimators are known to be well described, parameterised, and tested. Furthermore, labelling studies were carried out to assess the quality of data and correct it for labelling errors and inconsistencies. As far as our analysis goes, the existence of minimal mislabelled events across the final datasets (Table 2) suggests that label related problems should not be considered a major factor limiting the performance of our models. On the other hand, molecular descriptor studies confirm the limitation of predicting three dimensional properties exclusively from two and lower dimensional structural representations (Table 3). Moreover, the variance in performance enabled by different representations not only reflected their influence on the success of each modelling task, but also suggests that there is scope for the improvement of relevant molecular descriptions.

These findings align well with the learning curve results (Figure 6), which revealed the scope for model performance improvement through the implementation of strategies designed to decrease the bias and variance of models.

In summary, we conclude that the incomplete representation of molecular structures may be a major factor restricting the predictive power of our models to higher performances. In the next chapter, we attempt to build higher performing models through the improvement of molecular descriptions. We implement a library of low dimensional features designed to characterise the propensity of intermolecular contacts of a given compound, as this seems to be relevant information to all the properties here considered.

2.4 Methods

Software and Databases

Crystal structure information was extracted from the Cambridge Structural Database (2020 release) using the CSD Python API (v.3.0.0).[103] Labels were calculated using our *in house* developed algorithms as described in the ***Chapter 2 / Methods / Labelling Algorithms*** section. Molecular representations were calculated using the RDKit chemoinformatics package (2019.09.3).[133] Data manipulation was handled by the Pandas (v.0.25.1) and NumPy (v.1.18.1) packages. Statistical methods and distance metrics were implemented using the Scipy (v.1.4.1) package. Dataset visualisations, machine learning classifiers, hyperparameter optimisation routines, model performance metrics, confusion matrices, learning curves were implemented using the Scikit-learn (v.0.22.1) and MatPlotlib (v.3.1.2) packages. Outlier detection methods were implemented using the Pyod (v.0.8.3) library. All implementations were executed in Jupyter Lab (v.1.1.4) using Python version v.3.7.

Data Collection

The Cambridge Structural Database (CSD) acted as our primary source of data. Filtered queries were carried out to ensure that each set of molecules retrieved held the relevant information from which labels and molecular representations could be accurately calculated.

In general, the CSD was searched for all organic crystal structures of a single chemical component, excluding any metals, salts, and ions, as these present additional challenges that are not addressed in this study.[61,91,134,135]

Additional filters were applied according to the label calculation requirements of each property. For the HBND dataset, entries with disorder, errors or incomplete information about crystal atomic coordinates or hydrogen bonds were discarded, as they would not provide enough information for accurate network dimensionality calculation. Similarly, entries with errors or incomplete information about crystal atomic coordinates were discarded from the crystal habit dataset. Finally, only entries with incomplete information about crystal atomic coordinates were discarded in the polymorphism case.

Labelling Algorithms

HBND. Label assignment was based on a modification to the method described in *Bryant et al.*[84]. CSD API methods were used to mimic hydrogen bond intermolecular interactions between multiple copies of the same molecule, producing two network expansions of different sizes. The square roots of the eigenvalues of the covariance matrix containing the atomic coordinates that resulted from each expansion were computed. The three largest values were used as estimates of the three spatial dimensions of each network. Finally, the ratios between corresponding dimensions of each network expansion allowed us to deduce the number of directions along which the network grew.

Habit. Crystal habit assignment was based on the Bravais-Friedel-Donnay-Harker (BFDH) morphology.[95,96] The CSD API implementation of the BFDH method[136] was used to simulate a crystal for each molecule in the dataset and return the magnitude of its span across each of the three dimensions of space. From those, any molecule with at least one measure size equal to 0 (due to round ups of three decimal points) or above 100 (extreme outliers) were discarded. The three main dimensions of a simulated crystal in decreasing order of magnitude were defined as length (l), width (w), and depth (d). Finally, we developed a set of rules based on the ratios between these dimensions to separate crystals into distinct classes of morphology. Label assignment was carried out as follows:

$$habit\ (l,w,d) = \begin{cases} plate, & \frac{l}{w} < \frac{w}{d} - 1 \\ needle, & \frac{l}{w} < \frac{w}{d} + 1 \\ block, & \frac{w}{d} - 0.5 < \frac{l}{w} < \frac{w}{d} + 0.5 \end{cases}$$

The constants 1 and 0.5 were implemented to allow margins that promote a good separation of classes. All datapoints not classified by this rule were discarded from the training set for being considered ambiguous and potentially impose additional challenges to the model learning procedure. Edge cases like crystals landing near margins were visually inspected to validate the above set of rules and tune its constants. A visual representation of this labelling algorithm applied to the crystal habit dataset can be found in ***Supplementary Information C***.

Polymorphism. Labels were assigned using assumptions similar to those proposed by *Cruz-Cabeza et al.*[131] In the CSD, molecules with a polymorphic nature are tagged. The polymorphic subset included all CSD polymorphic-tagged molecules for which at least two

different crystal forms are present in the database. The monomorphic subset included all CSD non-tagged molecules with *at least three* crystal structure submissions to the database, but for which no multiple packing forms were yet identified. Under this scenario we assume that molecules have been well studied, such that if a different form was to exist, it would have already been found.

Labelling Studies

In CSD terminology, an *entry* corresponds to a crystal structure submitted to the database and a *family* corresponds to the collection of entries based with exactly the same molecular composition. Note that crystal structures based on the same molecule, but with different solvents, changes in composition or different rations of components correspond to entirely new families.

HBND & Habit. The CSD query output was aggregated by families, which were then classified using a rule-based system. Families were first split based on whether they have a polymorphic tag, then based on their number of associated entries, and finally according to whether label values across resubmissions of each crystal structure were consistent. The labelling error estimation was based on the subset of polymorphic tagged families that possess multiple resubmissions for each of its crystal forms. The estimate was calculated as the fraction of those whose labels vary across the same crystal form. The labelling stability estimation was based on the subset of families with multiple crystal structure submissions. The estimate corresponds to the fraction of those to which more than one label value was assigned across all their corresponding entries.

Polymorphism. The monomorphic assumption error estimation was based on the subset of polymorphic tagged families whose entries were submitted to the database at three or more different points in time. The estimate was calculated as the fraction of those

that only received a new crystal structure form in their third (or later) submission event. Submission events correspond to different dates at which new entries of a given family are added to the database.

Data Processing

Data processing included cleansing, scaling, outlier handling, and class size balancing. During cleansing, entries containing molecular descriptors with infinite or missing values were removed because they cannot be handled by the statistical methods to be used.

Domain outliers were also discarded because they correspond to structural representations that cannot physically exist. Domain outliers are identified by comparing the feature values of each molecule to the corresponding upper and lower bound values that the feature could realistically assume in the physical world. All remaining entries were scaled to zero median and unit interquartile range, promoting comparable feature contributions in classifiers that operate based on distance metrics between data points. At this point, the dataset was treated for extreme outliers. The extreme outlier discarding method examines each feature independently and discards points whose scaled feature values – one or more – are above the extreme threshold or below its symmetric value. Extreme outlier detection must only occur after dataset scaling, such that all features are on comparable scales. In this setting, a single extreme value threshold is required for the identification and filtering of extreme outliers. The threshold has been set to 100 throughout the entire project. Accordingly, any example containing at least one scaled feature above 100 or below -100 was discarded, since those start to differ significantly from the majority of observations. Domain and extreme outliers are formally introduced and described in ***Chapter 4 / 4.1.5 The outlier detection problem***. The scaler's motivation, as well as the extreme outlier threshold optimisation and selection can be found in ***Supplementary Information D***.

Class size balance was achieved by an under-sampling procedure, which consists of reducing over-represented classes to the size of the smallest class by random sampling. A class balancing report that allows comparison of the distribution of a given descriptor across different datasets was implemented to check that each class sample is representative of the pool that it was drawn from.

Molecular Representation Studies

High-, mid- and low- molecular representations were calculated for all molecules of the dataset. Features were processed as described in ***Chapter 2 / Methods / Data Processing***.

Given that no prior knowledge is available about the framework that best suits the modelling task, the performance of each molecular representation was tested across three different model estimators: a linear, a nonlinear and an ensemble estimator. These included a linear kernel SVM, an RBF kernel SVM, and a Random Forest. Each estimator was fitted to the training set using a hyperparameter grid search under a 3-fold cross validation routine. Optimised models were evaluated as the mean accuracy and F1 score of the cross validation. The performance of each molecular description was then measured by the mean test accuracy and mean test F1 scores obtained across the optimised estimators.

Dataset Visualisation

The t-Distributed Stochastic Neighbour Embedding method has two main hyperparameters, which greatly influence the final dataset visualisation. Perplexity is responsible for the balance between conserving the local and global structure of data, whilst the learning rate controls the step size of the optimisation procedure. Different

hyperparameter values were tested via grid search, and although the arrangement of points varies between projections, the general effect and overall conclusion are consistent.

The method was implemented with a perplexity of 40 and a learning rate of 10, which lies within the limits of 5 and 50 recommended by *Hinton et al*.[132] Results were visualised under a colour scheme matching points to the class they belong to.

Machine Learning Models

The processed dataset was divided into a training (70%) and test set (30%). Several statistical methods were considered and a full list can be found in the ***Supplementary Information B.*** Each statistical model was subject to a grid search hyperparameter optimisation under a 5-fold cross validation routine on the training set. Mean cross validation accuracies and F1 scores were used to compare and select model frameworks.

The selected estimator and hyperparameters were refitted to the entire training data. Model performance measures such as test set accuracies, F1 scores, confusion matrices, and learning curves were computed. The model was also evaluated for a random test set scenario based on y-scrambling,[137] which randomly shuffles the labels of the test examples. The prediction accuracy on this new set was held as the baseline model performance and any prediction accuracies above that were considered better than random.

Learning curves were computed from the entire class-balanced dataset subject to a 5-fold cross validation, using hyperparameter optimised estimators.

Chapter 3 – Improving chemical datasets

In Chapter 3, we address the challenges faced during chemical dataset construction and test whether there is room for the improvement of molecular structure description. We explore concepts such as molecular feature design, redundancy, and relevance. This exercise confirmed that improving structural descriptions through molecular feature design is a powerful but challenging process that requires a continuous effort in order to generate significant QSPR results. The key outcomes include a library of novel low dimensional features aimed to describe the intermolecular interaction propensity of molecular structures.

3.1 Introduction

3.1.1 Molecular descriptor design

Chemoinformatics approaches are able to model complex systems from large volumes of experimental data, with no explicit understanding or consideration of the underlying physical processes.[1,138] In a QSPR context, such systems correspond to molecular structure – property relationships, and structures are often represented as collections of molecular descriptors. These molecular representations are expected to consider descriptors that cover all of the key structural aspects required to accurately describe the property being modelled. Unfortunately, QSPR studies focus on structure – property relationships that are not well understood and identifying the complete set of relevant descriptors can be a tough task.

Each molecular descriptor is calculated by a mathematical procedure that compresses chemical information into a single value aiming to represent and describe a very specific aspect of the molecular structure.[139] The extensive literature on molecular feature

design shows a remarkable community effort towards the complete description of molecular structures. The number of structural features is large and keeps increasing.[140] A review of molecular descriptors can be found in *Khan et al*.[111] In addition to simple atoms, bonds, functional groups, and fragment counts, two and lower dimensional descriptors also cover aspects related to the van der Waals surface,[141] electrotopological state indices,[142] partial charges,[143] molecular shape,[144] branching,[145] polar surface area,[146] various physicochemical attributes[147] and others. Unfortunately, low dimensional molecular descriptors are often correlated with each other and whilst certain aspects of the molecular structure are well characterised, others seem to remain uncaptured. The incomplete description of the molecular structure of compounds is a recurrent issue that often compromises the quality of QSPR modelling results.

Feature engineering is the process of creating features that enhance the predictive capability of a machine learning model. New features are often built with the purpose of explicitly providing important information that is already contained in the original ones, or to add new relevant information that contributes to a more complete description of the target variable. The latter is referred as feature design and it can be summarised into a three-step process. First, domain knowledge is required to define the type of information that would be useful to provide to the mapping task. Then, an algorithm is implemented to capture that information and encapsulate it into a meaningful value. Finally, the feature must be tested and possibly optimised.

Improving the diversity of molecular descriptors and information content of structural representations is key to increase the predictive power of chemoinformatics models. Indeed, the most suitable and powerful solution involves the thoughtful design of new and relevant molecular descriptors. The task is greatly open to creativity, but must follow some design rules and ensure the chemical accuracy of the output.[110] For example, a

descriptor must be invariant with respect to the labelling, order and numbering of atoms, and it must possess an unambiguous algorithmically computable definition. Furthermore, descriptors should also have a clear structural interpretation and no trivial correlations with other existing ones.

3.1.2 Intermolecular interaction propensity

Two major works helped to set the basis for understanding intermolecular interactions at a quantitative level. *Abraham et al.* [148] was first to formulate and report hydrogen bond constants as a formal way to measure the propensity of functional groups to act as hydrogen bond acids and bases. Later on, *Hunter et al.* [149] reported a scale to reflect the intermolecular interaction propensity between two molecules, based on their functional group content and corresponding hydrogen bond constants.

Both studies embrace the idea that intermolecular interactions can be generally considered of hydrogen bond nature, even if the interacting regions do not necessarily correspond to conventional hydrogen bonding sites. Free energy is the key observable from which such events can be understood and formalised.[148,149] For a given temperature, the free energy of an intermolecular interaction is governed by two main factors: the unfavourable change in entropy of bringing two molecules together and the favourable change in enthalpy associated with the formation of a new stable contact.[150–152] The basic theory of intermolecular interactions further separates enthalpy into four components: repulsion, induction, dispersion and electrostatic. Experimental evidence shows that repulsion, induction and dispersion contribute negligibly to explain intermolecular interactions, suggesting that such contacts are predominantly guided by electrostatics.[148,153–155]

The dominant electrostatic interactions between two molecules generally occur between the sites of maximum and minimum electrostatic potential energy of the van der

Waals surface of the interacting molecules.[149] Furthermore, maxima are often located near a hydrogen atom and minima usually occur over a lone pair of electrons or close to a π-electron density region.[149] Hence, intermolecular interactions may generally be considered of hydrogen-bonding nature. A schematic representation can be found in **Figure 7**.

The combined efforts of *Abraham et al* [148] and *Hunter at al* [149] led to the development of a thermodynamic ranking, where the intermolecular interaction propensity between two molecules is based on the hydrogen bond acidity and basicity like character of their functional groups, each defined by the hydrogen bonding constants α and β.

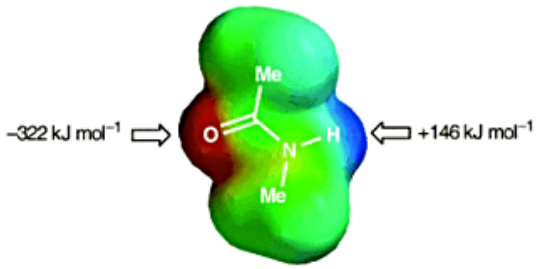


Figure 7 Representation of the electrostatic potential energy plotted on the van der Waals surface of N-methyl acetamide. Positive regions are shown in blue, negative regions are shown in red, and green regions are neutral. The two primary hydrogen-bonding sites of the molecule are the NH group and the carbonyl oxygen atom, which indeed correspond to the regions of maximum and minimum electrostatic potential (adapted from *Hunter et al.*[149]).

3.1.3 The assessment of feature usefulness

Feature usefulness refers to the value of a feature within a given context. Feature redundancy measures the unique information content of a feature within a dataset, whilst feature relevance evaluates it with respect to a modelling task.[80] Accordingly, feature redundancy assessments are often based on the level of association amongst explanatory features, which is usually characterised in terms of correlation or mutual information.[156,157]

On the other hand, feature relevance assessments are often based on the level of association between explanatory features and the target variable, which may be best characterised in terms of model performance.[156,158] In both cases, higher association values correspond to higher degrees of feature redundancy or relevance.

In *Chapter 5*, we revisit the concepts of feature redundancy and relevance to elaborate on the assessment of feature usefulness and propose a feature selection pipeline within the minimum-redundancy maximum-relevance framework. Refer to ***Chapter 5 / 5.1.2 The feature selection process*** for more details on feature redundancy and relevance assessment procedures and limitations.

3.1.4 Wrap up

In the last chapter we investigated whether it was possible to predict material properties using information derived from their constituent molecules. The limited predictive power of the resulting models was found to be tightly related to the insufficient relevant information content within the features of the corresponding chemical datasets.

In this section we address these faults and test whether there is room for the improvement of low dimensional molecular structure representations. Here, we implement a library of low dimensional features designed to capture the intermolecular interaction propensity of a given compound, as to our knowledge no open source molecular feature captures such aspect of molecular structures. The usefulness of each feature is investigated in light of each case study, to assess whether it adds new information to the dataset and if any information added is relevant to the prediction task.

3.2 Results

3.2.1 Feature library implementation

Building on the parameterisations made by *Abraham et al.* [148] and *Hunter et al*.,[149] we developed a chemoinformatics library of molecular features that attempt to capture the intermolecular interaction propensity of different molecular structures.

Molecular descriptors are originally designed for various different target properties or biological activity. Therefore, describing the potential for self-self intermolecular interactions might not have been considered important for previous QSPR/QSAR studies. However, most solid-state material properties are greatly influenced by the arrangement of their molecular constituents in the crystal structure, which is fundamentally guided by intermolecular contacts.[159] Providing a way to consider the intermolecular interaction propensity of molecular structures in QSPR modelling may indeed be of great benefit to a wide variety of Chemoinformatics problems, including the prediction of hydrogen bond network dimensionality, crystal habit, polymorphism, and many others. Hence, the implementation of the present library was motivated by the lack of commercially or freely available features describing the intermolecular interaction propensity of molecular structures, and the existence of extensive preliminary work on the subject, which makes it possible to capture such information in a small collection of low dimensional and highly interpretable features.

The algorithm implemented here calculates all intermolecular interaction propensity (IIP) features in three simple steps. First, each molecule is searched for its functional group content. The hydrogen bonding constants α and β [148,149] of each identified group are retrieved, and used for the computation of the 24 IIP features in the library. Half of these are based on the constant α and attempt to describe the hydrogen bond donor like character of the molecule, whilst the other half is based on the β constant and attempt to describe its

hydrogen bond acceptor like character. All feature names and meanings are listed in **Table 5**. These include the sum, average, minimum, and maximum hydrogen bond acidity and basicity like character values across all functional groups within a molecule, which aim to summarise the absolute intermolecular interaction propensity of molecules. Others include the sum of α and β values normalised by molecular weight, log P, or van der Waals surface area, and a few others inspired by *Crippen et al.*[147] and *Labute et al.*[141] These aim to capture the global and local distribution of the hydrogen bond acidity and basicity-like character across a molecular structures, respectively.

Table 5 Intermolecular interaction propensity feature names and definitions. The α_VSA* and β_VSA* are descriptors based on an approximate accessible van der Waals surface area calculation for each atom v_i along with some other property p_i. Each descriptor in a series is defined to be the sum of the v_i overall all atoms i such that p_i is in a specified range, normalised by the total surface area v. Accordingly, α_VSA* represents the fraction of van der Waals surface area with a hydrogen bond acidity like character (α) between 0 and 1 (α_VSA0), 1 and 2 (α_VSA1), 2 and 3 (α_VSA2), 3 and 4 (α_VSA3), and above 4 (α_VSA4). On the other hand, β_VSA* represent the fraction of van der Waals surface area with a hydrogen bond basicity like character (β) between 0 and 2 (β _VSA0), 2 and 4 (β _VSA1), 4 and 6 (β _VSA2), 6 and 8 (β _VSA3), and above 8 (β_VSA4). The ranges were determined by a percentile subdivision over a large collection of compounds.

IIP Feature	Definition
max_ α, max_β	Maximum hydrogen bond acidity and basicity like character across all functional groups within a molecule.
min_ α, min_β	Minimum hydrogen bond acidity and basicity like character across all functional groups within a molecule.
total_ α, total_β	Sum of the hydrogen bond acidity and basicity like character of all functional groups within a molecule.
mean_ α, mean_β	Average hydrogen bond acidity and basicity like character of all functional groups within a molecule.
α_mw_norm, β_mw_norm	Sum of the hydrogen bond acidity and basicity like character of all functional groups within a molecule, normalised by total Molecular Weight.
α_logp_norm, β_logp_norm	Sum of the hydrogen bond acidity and basicity like character of all functional groups within a molecule, normalised by total log P.
α_vsa_norm, β_vsa_norm	Sum of the hydrogen bond acidity and basicity like character of all functional groups within a molecule, normalised by total van der Waals surface area.
α_VSA*, β_VSA*	MOE-type descriptors using α or β values and surface area contributions, *i.e.* fraction of the accessible van der Waals surface area with a hydrogen bond acidity / basicity like character within particular ranges.

The current library implementation is able to identify all 46 different functional groups listed in the original paper for which α and β values are reported (***Supplementary Information E***).[149] We anticipate that this library must be used with care, given that no other groups will be accounted for, during feature calculation. However, note that the algorithm supports the easy extension of the functional group list, as long as α and β values are computed beforehand. The calculation of both constants can be achieved as described in the original literature.[148,149] The implementation of this library of descriptors was a collaboration with the CCDC partners in this project.

3.2.2 Feature usefulness assessment

3.2.2.1 Feature redundancy

We test the redundancy character of each IIP features within the context of each dataset – HBND, crystal habit, and polymorphism - by evaluating how well the values of a particular IIP descriptor can be predicted using only features from each original dataset. If a new descriptor can be accurately predicted from the original explanatory variables, then the information content provided by the descriptor can also be deduced from them. In contrast to some feature redundancy methods, this experiment considers feature dependencies and the regressor version of the original modelling framework (estimator and hyperparameters), such that redundancies missed at this point would also not be identified during the modelling of our material properties. The adjusted coefficient of determination[160] (R^2) achieved for each test was used as a measure of feature redundancy, and lower values are associated with a stronger uniqueness character. Note that a feature is said to be unique within a particular dataset if most of the information it carries cannot be provided by any other descriptor in the dataset. Furthermore, unique features greatly contribute to the increase of the information content of the datasets they belong to.

Table 6 provides the key for the conversion of R^2 values into feature redundancy labels and reports the number of IIP features classified as having a high, moderate, and low redundancy character within each original dataset. The library was found to contain features from the entire spectrum of redundancy. For example, *α_mw_norm* and *β_mw_norm* seem to be quite redundant (high R^2 value), whilst *α_logp_norm* or *β_logp_norm* appear to be more unique (low R^2 value) across all three datasets. See ***Supplementary Information F*** for more details.

Table 6 IIP feature redundancy results for the crystallizability case study. The table below contains the conversion of R^2 values into feature redundancy labels. The number of IIP features that fall into each feature redundancy category is reported for each case study. Feature redundancy was measured using the adjusted coefficient of determination (R^2), which represents the proportion of variance of a continuous dependent variable that the independent counterparts can explain under a given statistical modelling framework.[161] Its adjustment accounts for its inflation when the number of explanatory variables increases.[160] R^2 is one of the only popular and intuitive measures for regression problems that enables performance scores within an intuitive and bounded scale, allowing us to categorise and compare features according to their level of redundancy. R^2 values were obtained from the modelling of each IIP feature using the original data. See ***Chapter 3 / Methods / Feature Redundancy Studies*** for details.

R^2 value	Feature redundancy	HBND	Habit	Polymorphism
1.00 - 0.75	High	4	9	1
0.75 – 0.50	High to moderate	14	10	11
0.50 – 0.25	Moderate to low	4	1	6
0.25 – 0.00	Low	2	4	6

3.2.2.2 Feature relevance

A feature is said to be relevant if it directly or indirectly highlights structural aspects that are of relevance to the property prediction task. In practice, relevant features often contribute to the increase in model performance, especially if these also have a low redundancy character. Useful features that are not unique rarely lead to an increase in model

performance, because the useful information that they add to the modelling task is already being provided by other existing features. For this reason, we only conduct feature relevance studies on IIP features that were found to have a moderate to low redundancy character on each particular case study, *i.e.* those with an R^2 below 0.5 (see ***Chapter 3 / 3.2.2.1 Feature redundancy*** for details).

We assess feature relevance by evaluating the increase in model performance enabled by adding each individual IIP feature to the original dataset, during the property modelling task. Feature relevance was measured as the difference in performance of models built from the original datasets and those built from original feature plus an extra IIP feature each time.

Table 7 shows the number of unique IIP features that were also found to be useful to each prediction task. Only features yielding a statistically significant model performance increase were considered to be relevant. The maximum individual contribution is also reported. The IIP library was found to contain useful features for the prediction of crystal habit. The *β_logp_norm* feature alone contributed to an increase of 0.5% of the overall model performance, whilst the *α_logp_norm* had an individual contribution of 0.4%. See ***Supplementary Information G*** for more details.

Table 7 IIP feature relevance results for the crystallizability case study. Note that only unique features were considered for relevance studies. The number of unique features accounts for those assigned to the moderate to low and low redundancy categories, *i.e.* $R^2 < 0.50$ (**Table 6**). Feature relevance was measured as the difference in model performance (mean F1 score of five 2-fold cross validation routines) between a reference model built from each original dataset, and those built from the original dataset plus an extra IIP feature each time. The statistical significance of the differences in performance between reference and the remaining models was evaluated with a Wilcoxon sign-ranked test and a two-tail p-value cut-off of 0.05. The maximum unique contribution reflects the maximum individual model performance increase yielded amongst all useful IIP features for each case study. See ***Chapter 3 / Methods / Feature Relevance*** **Studies** for details.

	Nr Unique Features	**Nr Relevant Features**	**Max unique contribution**
HBND	6	0	NA
Habit	5	2	0.5%
Polymorphism	12	0	NA

Note that these feature relevance studies have two main limitations. Firstly, they require the assessment of model performance on multiple training and test sets with no overlap of observations, to guarantee the independence of data samples and in turn the validity of statistical hypothesis test results. This is particularly challenging to implement for small datasets, as it often leads to the increase of the variance and decrease in performance of each individual model produced. Furthermore, nonparametric statistical hypothesis significance tests have less power than other parametric alternatives. However, it was possible to achieve sufficient feature relevance assessment results and draw conclusions about the process of molecular feature design as means to improve the predictive power contained in the features of QSPR datasets.

3.3 Conclusion

The successful QSPR modelling of a given material property requires the consideration of the factors that are relevant to its description. Molecular representations

are designed to account for all those aspects, but the incomplete understanding of the chemical system and/or the limited pool of available molecular features often restrict the level of description that can be achieved. This is a recurrent issue that often compromises the quality of QSPR results.

In line with the findings of Chapter 2, we believe that improving the diversity of molecular features is a powerful and viable solution to increase the information content of structural representations and in turn the predictive power of QSPR models.

Expanding on the work conducted by *Abraham et al* [148] and *Hunter at al* [149], we develop and report a library of 24 low dimensional and highly interpretable molecular features that attempt to describe the intermolecular interaction propensity of molecular structures. We further tested each new feature for its redundancy within a particular dataset and relevance for a given property prediction task.

The library contained unique features that were shown to increase the information content of structural representations, and useful features that led to the increase of model performances. These findings suggest that our library not only contributes to the description of structural aspects that no other tested low dimensional feature does, but also that the new information captured seems to be of relevance for the modelling of different crystal and solid-state material properties. More broadly, our findings illustrate the challenges of adopting feature engineering and molecular feature design strategies for the improvement of QSPR results. The small increase in model performance suggests the scope for the improvement of the predictive power of chemical datasets through molecular feature design, but it also reveals the amount of domain knowledge and effort required to make it an effective solution. Indeed, this exercise confirms that improving structural descriptions through molecular feature design is a powerful, but challenging process that requires a continuous effort in order to generate significant results. Although conclusions are derived

from models where information about intermolecular interaction and energy seems to be important, we speculate that, barring further developments in the description of molecular structures, most QSPR models will face limitations in their predictive power and fail to meet the requirements for reliable practical use.

3.4 Methods

Feature Library Implementation

The IIP library depends upon a series of RDKit methods. The first step neutralises the charges of all input molecules to guarantee that functional group representations are standardised for efficient identification. Groups are identified during a sub structural search, guided by SMARTS pattern representations of each functional group. Group matches are adjusted in cases where small functional groups are found within bigger groups. For example, a molecule with a carboxylic group would also produce a match to an alcohol group, and the match adjusting step corrects for these scenarios. Finally, the α and β constants of all adjusted matches are retrieved and used for descriptor calculation. Molecular weight, van der Waals surface area (VSA)[141] and logP[147] are normalisation factors also calculated using RDKit methods.

Feature Redundancy Studies

The values of all 24 IIP descriptors were calculated and merged onto unprocessed versions of the original datasets (see ***Chapter 2 / 2.2.2 The datasets***). Extended datasets were processed as described in ***Chapter 2 / Methods / Data Processing***.

Original features were used to model each of the 24 IIP descriptors. Statistical frameworks equivalent to those initially used for property prediction were adopted to ensure that feature associations not found here would also not be found during material property

modelling. Note that IIP descriptors are continuous variables, meaning that the statistical frameworks adopted in these studies were based on the Support Vector Regressor method from scikit-learn. Hyperparameters were set as previously reported (see ***Chapter 2 / 2.2.3 The models*** for reference).

Adjusted coefficient of determination (R^2) has been previously adopted as a performance measure for support vector regressors using various kernels.[162–164] R^2 represents the proportion of variance of a continuous dependent variable that the independent counterparts can explain under a given statistical modelling framework.[161] Mean R^2 values were calculated for each IIP descriptor over a 3-fold cross validation routine. R^2 values were converted into feature redundancy categories according to the following key: low ($R^2 < 0.25$), low to moderate ($0.25 < R^2 < 0.50$), moderate to high ($0.50 < R^2 < 0.75$), and high ($0.75 < R^2$). Finally, the number of IIP features in each category was reported for the three case studies of this project.

Feature Relevance Studies

The values of all 24 IIP descriptors were calculated and merged onto unprocessed versions of the original datasets (see ***Chapter 2 / 2.2.2 The datasets***). Extended datasets were processed as described in ***Chapter 2 / Methods / Data Processing***.

A collection of datasets including all original features plus a different extra IIP feature each time were produced. Only the datasets with IIP features that were found to have a moderate to low redundancy character within each original dataset, *i.e.* $R^2 < 0.5$, were further considered (***Supplementary Information F***).

Statistical significance tests for comparing machine learning algorithms are often based on 5 repeats of 2 cross validation routines. Two folds are chosen to guarantee that each observation appears only once in the train or test sets for each model skill estimation.

This strategy ensures the independence of data samples from where each performance value is estimated, given that this is a common assumption amongst statistical hypothesis significance tests.[165] Nonparametric paired tests such as the Wilcoxon signed-rank test[166] make fewer assumptions about the distribution of the model skill. These seems the best option given that only 10 performance values are retrieved for each model.[167]

According to the approach described above, datasets were divided into 5 equally sized samples. Model performances were measured as the mean F1 score value of five 2-fold cross validation routines based on the original statistical frameworks initially used for property prediction (see ***Chapter 2 / 2.2.3 The models*** for reference).

Feature relevance was computed as the difference between model performances when only original features are considered (baseline reference) and when IIP descriptors are considered one at a time. The statistical significance of the difference in performances was evaluated using the nonparametric Wilcoxon signed-rank test and a two-tail significance value cut-off of 0.05, under the null hypothesis that no difference exists between the performances of each pair of models. Useful features were counted and reported for each case study, along with the value of the maximum individual feature relevance achieved amongst those.

Chapter 4 – Increasing the practical value of models

In Chapter 4, we acknowledge the lack of relevant information content of molecular representations used for the prediction of our case studies, and design strategies to increase the practical value of limited performing models. Here, we cover the concepts of stationarity assumption and applicability domain in detail. Key outcomes include the generalisability and confidence tools designed to restrict models to inputs that can be predicted reliably and with confidence. The proposed strategies demonstrate that controlling the quality of inputs and outputs can help to identify the regions of chemical space where QSPR models are able to add great practical value, including those of limited performance.

4.1 Introduction

4.1.1 The vastness of chemical space

"Chemical space" is often defined as the span of all possible and imaginable chemical compounds.[168] Chemical space is incredibly vast, sparse, and diverse.[168–173] For example, the subset of small organic molecules can be defined to include all the compounds with a size up to 17 heavy atoms (carbon, nitrogen, oxygen, sulphur, and halogens).[174–176] Considering the connectivity patterns and stereochemistry allowed, this region alone is estimated to contain more than 166 billion viable molecular arrangements.[176]

Accordingly, the size of chemical space represents a challenge for its navigation and exploration.[177,178] To make it feasible, the chemical space is fragmented into smaller regions represented by diverse collections of molecules - virtual datasets and physical libraries. This process often entails i) the definition of chemical space in terms of features of interest, such as molecular weight or solubility, ii) the delimitation of a subspace by constraining those

features to certain value ranges, and iii) the rational sampling of molecules from the resulting chemical space region.[179]

In drug development, chemical datasets are often built for the virtual screening and identification of lead and optimised compounds.[180] With factors like size and heterogeneity playing a key role in increasing the chance of finding a high-quality hit, it is desirable for such datasets to cover as much of the chemical space as possible.[181,182] Accordingly, the sampling of chemical space regions is often guided by the concept of chemical diversity that encourages the production of representative, diverse, and manageable datasets.[183–185]

4.1.2 The stationarity assumption of statistical learning

Models can be assessed internally or externally. Internal model assessments provide estimates for the predictive power capacity enabled by a given training set. These refer to the maximum performance that the model could ever achieve using the modelling framework selected and the data at hand. This can be calculated using validation sets or using cross validation routines.[4] On the other hand, external model assessments provide estimates for how well the model is able to generalise to unseen examples, and the generalisability error is defined as the difference between internal and external model performances.[186,187]

The statistical learning theory framework assumes that past and future observations must follow the same data distribution because they are all produced by the same data generation process.[188,189] Under this assumption, highly predictive models can be expected to generalise well, *i.e.* retain the predictive power exhibited during internal assessment, across any new examples that comply with the distributional characteristics of the training set.[188,189]

During model construction, training and test sets are often produced by the random sampling of the same initial pool of examples. The distributional agreement between training

and test data complies with the stationarity assumption, allowing models to generalise well to unseen test examples and achieve the performance levels expected. However, this important assumption is generally not fulfilled in real world scenarios. After model deployment there is often little to no consideration about the distribution of the data to be evaluated. In these contexts, models tend to underperform and the extent to which the stationarity assumption is violated is tightly related to the loss of model generalisability and consequent decrease in performance.[66] This does not mean the model is no longer useful, but that the inputs are inadequate. Several strategies have been developed to maximise the generalisability scope of machine learning models, but these still do not provide clear bounds to their generalisation capacity.[190,191]

Many promising published models have been reported that do not perform according to expectations after deployment.[59–63] Understanding and defining the generalisability properties of machine learning models is required, but classic generalisability theory approaches like the Vapnik-Chervonenkis dimension[192], the Rademacher complexity[193], or the PAC-Bayes bounds[194] are hard to compute, mostly descriptive, and fail to provide useful generalisation bounds to a model in realistic contexts.[195,196]

In chemoinformatics, the set of molecules used to train a QSPR model is amazingly small when compared to the entire chemical space.[173] Despite the adoption of strategies like chemical diversity, the molecules of a given dataset still follow a particular distribution that differs from datasets representing other chemical space regions. Hence, no model generated from a finite set of molecules can be expected to generalise across the whole chemical space with the same predictive power.[197] Models must be used within their applicability domain in order to generalise well to unseen examples and preserve the high confidence of their predictions. The need for a strategy that formally identifies the applicability domain of a predictive model has fomented a great academic and industrial effort.

4.1.3 The applicability domain of data driven models

The applicability domain (AD) refers to the regions of space within which a given model can be used reliably, *i.e.* operate at its expected level of performance and provide predictions that can be trusted on an individual basis.[198] A well-defined AD makes QSPR models suitable for the reliable scanning of large virtual libraries, but also for the trustworthy assessment of individual molecules. This appealing feature soon became a requirement for predictive models, established by the Organization of Economic Co-operation and Development (OECD).[198]

Defining the AD of a model is a manifold problem and several strategies have been proposed over the years to cover its different dimensions. A schematic representation of the main dimensions of the AD is presented in **Figure 8**. The description, comparison, and scope of the main approaches are extensively described across the literature.[7,75,79,199–214] Ultimately, the reliable use of models can be achieved if all inputs that are likely to lead to prediction errors can be identified and excluded from the classification task. These can be grouped into three types of unreliable use cases, which set the basis for the different ways that the boundaries of the applicability domain can be defined.

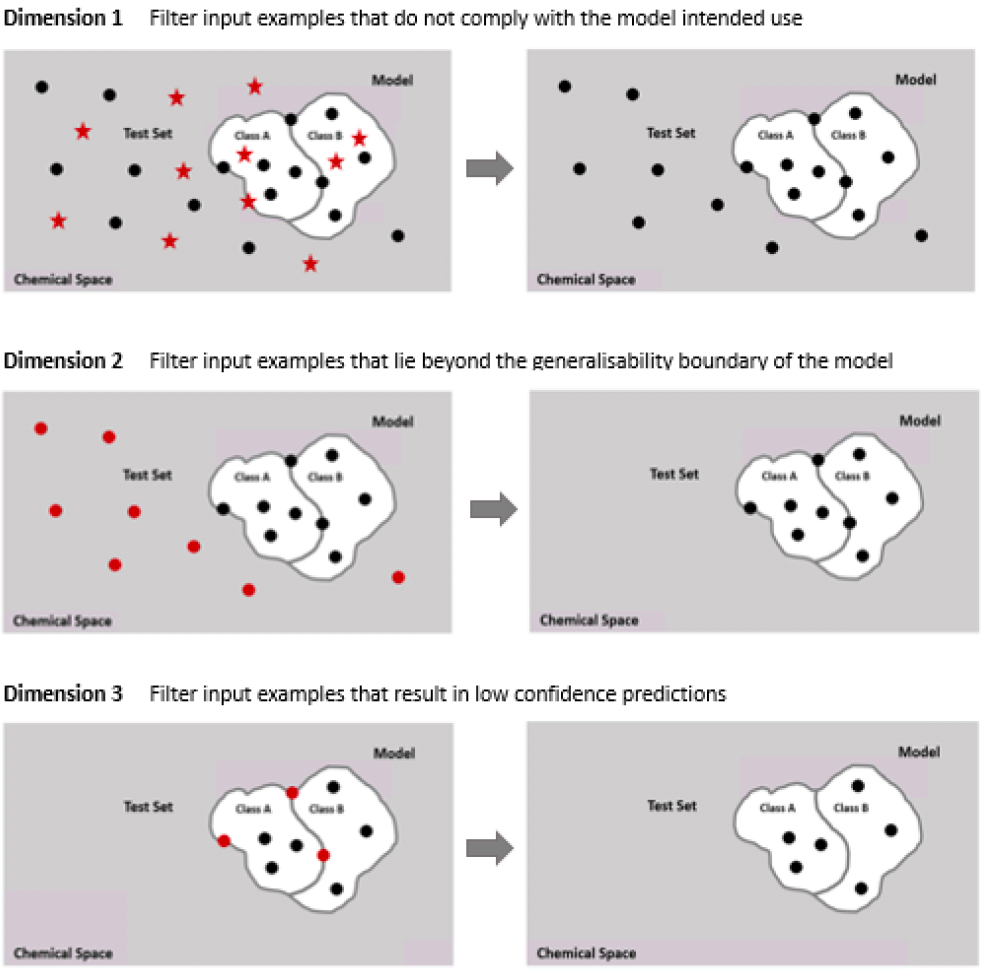


Figure 8 The main dimensions of a model's applicability domain. Dimension 1 filters input examples based on the specifications of the model. Dimension 2 filters input examples according to an outlier detection framework. Dimension 3 filters input examples based on prediction confidence estimates, using probabilities of class membership. Circle and stars represent test examples. The white area corresponds to the region of chemical space that the model focuses on and its limits correspond to the boundaries of its generalisation capability. The line across the white area corresponds to the boundary between classes.

The first approach is concerned with whether the input complies with the intended use of the model (**Figure 8, Dimension 1**). The validity of each use case is determined by assessing if its characteristics meet the model specifications.[81] In other words, new test examples must fulfil any requirements adopted during the construction of the training set. For example, a QSPR model built from a collection of organic compounds is only prepared to

predict organic inputs. Due to context dependency, the user has the responsibility to manually ensure that such requirements are met, but efforts towards the automatic filtering of inputs at this level are starting to happen.[82]

The second approach is concerned with the extent to which the model is able to generalise to new inputs (**Figure 8, Dimension 2**). Here, the scope of the applicability domain is determined by the structure and content of the training data.[75,215] According to the stationary assumption, models are expected to generalise to new inputs that follow its training set distribution. Thus, points that are far away from the training data or lying in low density regions of space tend to correspond to less reliable inputs. In statistics, this assessment can be formulated as an anomaly or outlier detection task. Anomaly detection systems are often implemented as a metric to assess the anomaly character of individual input examples with respect to the training set, where larger scores tend to correspond to larger prediction errors. A number of suitable strategies have been extensively discussed across the literature.[79,216–218] Moreover, additional insights on anomaly detection can be found in ***Chapter 4 / 4.1.4 The outlier detection problem***.

The third and final approach is concerned with the level of confidence with which the model outputs individual predictions (**Figure 8, Dimension 3**). The confidence of a given prediction varies according to how much the characteristics of the input example agree with the data patterns that the model uses to assign it to a given class. Thus, points closer to the model's decision boundary or lying in regions where classes overlap are more difficult to assess and tend to lead to less confident predictions. The confidence scores of QSPR classifiers are often derived from probabilities of class membership.[219–223] These may directly reflect the trust of a particular prediction and larger scores often represent more confident outputs. A review of different confidence estimation methods can be found in *Mathea et al.*[79]

During QSPR model development and implementation, one should consider each of the applicability dimensions described above. Furthermore, *Hanser et al.*[81] stress the importance of using these approaches in tandem to ensure that future model inputs are efficiently narrowed down to the examples that belong to the model's applicability domain. Accordingly, only points that are found in the intersection of these subdomains can be converted into valid, reliable, and confident predictions.

The applicability domain is a recent and active area of research.[211] Despite theoretical advances, not many solutions are actually implemented by developers or adopted by users. Furthermore, the few open source options that exist [224–226] are built on very specific methods whose performance seems to be context dependent.[75–77]

4.1.4 The outlier detection problem

In statistics, an anomaly or outlier is a data point that differs significantly from other observations.[227] These occur in most real-world datasets and often stem from errors in data, or simply represent genuine but discardable abnormalities or interesting artefacts.[228] Given that the presence of outliers can have a serious impact in data analysis and statistical modelling results, outlier detection is a growing concern in many research areas and application domains. Therefore, many approaches and methodologies have been introduced to address the problem. *Chandola et al.*[217] and *Pimentel et al.*[218] are two of the numerous examples that provide thorough reviews on outlier detection techniques and applications.

In chemical datasets, each molecule is often described by a set of structural features, all of which are numerical and only allowed to take values within the domain that reflects their physical reality. In this project, we make a distinction between domain, extreme, and structural outliers. Domain outliers refer to instances whose feature values – one or more – lie outside of their physical domain (*e.g.* negative molecular weight). Such errors arise due to

bugs in descriptor calculation algorithms or when molecules break their underlying assumptions, and resulting values are no longer accurate representations of their structure.[63] Extreme (univariate-like) outliers refer to instances whose feature values – one or more – are extreme. Although the instance is described by a set of realistic feature values, at least one of those is uncommonly small or large. Accordingly, the point is deemed an unusual case within the overall sample. Note that features are individually studied and no considerations about their interdependencies are made. On the other hand, structural (multivariate-like) outliers are found in the *n*-dimensional space and may refer to perfectly reasonable molecular representations with no extreme feature values, that simply diverge from the overall pattern followed by other observations. Although the instance may correspond to a realistic molecular representation, the combination of feature values that describe the compound does not fit with the internal correlation structure of the sample.

Generally, anomaly detection methods are used to build models of normality around the content and distributional characteristics of the training set. These models are then adopted to assess how much each new unseen example deviates from the default distribution, outputting anomaly values that reflect the distance between both statistical objects. These scores are finally compared to a decision threshold and inputs are considered anomalous if the threshold is exceeded.[218] In the QSPR modelling context, such results would further suggest that those instances would be out of the model's applicability domain. Anomaly threshold values are often derived from the distribution of scores computed for all training examples, as the mean score plus x-times the standard deviation (parametric method) or simply as one of its quantiles (nonparametric method).[79] The appropriate choice of the threshold depends on the intended use case.[217]

Extreme and structural outlier detection follow the same strategies to those used for univariate and multivariate anomalies, meaning that they also inherit their disadvantages.

Outlier detection is not a trivial task and many challenges come with it, such as the subjective selection of anomaly thresholds,[79,229] but also the fact that performance of multivariate anomaly detection approaches seems context dependent[213,230–232] and often deteriorates in high-dimensional spaces.[229,233]

4.1.5 Wrap up

In the last chapter we investigated whether there was room for the improvement of low dimensional molecular structure representations. The exercise suggested that enhancing the predictive capability of limited performing QSPR datasets through molecular feature design is a challenging process that requires a substantial effort in order to generate significant results.

Our QSPR models may not perform up to most practical standards, yet their reasonably better-than-random results are enough evidence to support the existence of useful structure - property relationships in the data. In this section we attempt to increase the practical value of our (or any other) QSPR classifiers by restricting them to the regions of space where they can be used reliably.

First, we conduct chemical diversity studies to demonstrate that no QSPR model can be expected to generalise indefinitely across chemical space and motivate the importance of the applicability domain concept. Then, we propose two novel and complementary applicability domain identification strategies: the generalisability tool that aims to delimit the regions of chemical space where models hold their predictive power, and the confidence tool that aims to delimit the regions of chemical space where models produce confident predictions. **Figure 9** clarifies where each of these methods sit in the QSPR prediction pipeline. The performance of both tools is investigated in light of each case study.

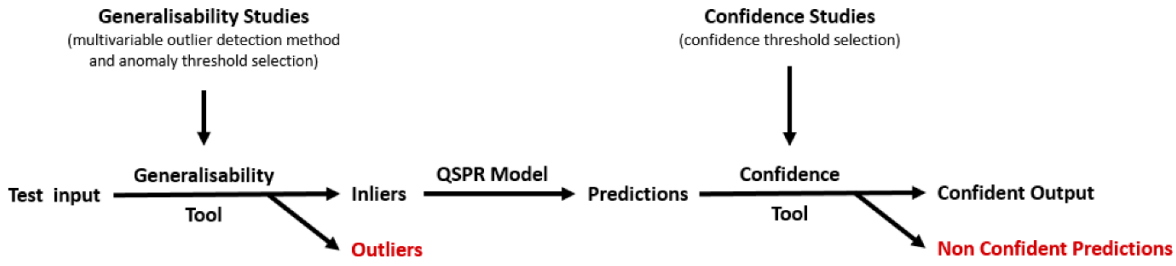


Figure 9 The generalisability and confidence tools within the QSPR prediction pipeline.

4.2 Results

4.2.1 Chemical diversity studies

In line with recent literature and following the above definitions of internal and external model evaluation, we assess the generalisation capability of a given model on test sets of increasing chemical diversity, to illustrate why QSPR models should not be expected to generalise indefinitely across chemical space.

According to the stationarity assumption of statistical learning theory, the model's ability to generalise to a given test set is dependent on the agreement of the training and test set distributions. Note that in this study, chemical diversity is a conceptual tool used to simulate the gradual increase of training and test set distributional differences. Diversity metrics mainly attempt to cover molecular composition, functionality, size, and shape.[185,234–236] Accordingly, molecular weight intuitively correlates to diversity factors, acting as a simple, general, and efficient proxy for chemical diversity in a drug development context.[3,81,83,236,237]

Here, the chemical diversity of a dataset was measured as the average molecular weight of its instances. Training and test sets were built based on molecular weight constraints, under the assumption that test sets with a higher average molecular weight are more chemically diverse. Moreover, larger differences between the chemical diversity values

of the training and test data should reflect larger distributional differences between the sets. We refer to this setup as a controlled model deployment scenario. The crystallizability dataset was used in this experiment. The training set was used to build a predictive model and assess its internal performance, whilst test sets were used for external model evaluation. The generalisability errors were calculated as the difference between internal and test [external] model performances. These results were plotted against the chemical diversity value of each corresponding test set.

Results are shown in **Figure 10**. The findings suggest that when the crystallizability predictive model is evaluated on different test sets, the increase of chemical diversity across these translates into an increase in the model's generalisability error. In other words, the decrease of model generalisation capability and test performance seems to be directly related to the extent of the differences between training and test set content.

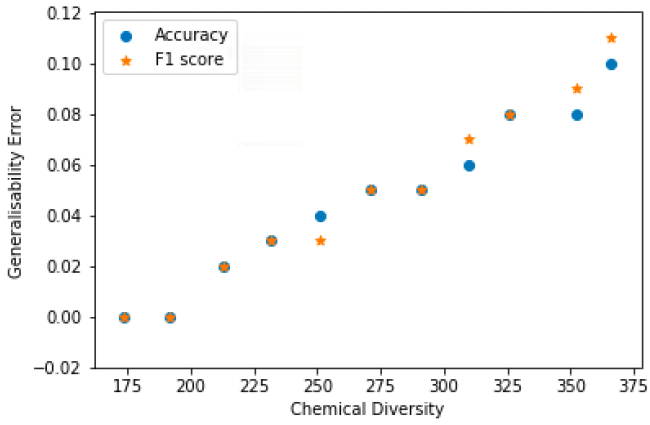


Figure 10 Chemical diversity study results for the crystallizability case study. The generalisability error plot shows the evolution of the model generalisability error on test sets of increasing chemical diversity. Training and test sets were produced from the crystallizability dataset as described in ***Chapter 4 / Methods / Crystallizability Dataset & Model***. Here, the chemical diversity of a set was measured as the average molecular weight of its instances. The chemical diversity study was then conducted as described in ***Chapter 4 / Methods / Chemical Diversity Studies***. The internal assessment of the model obtained from the crystallizability training set reported an overall accuracy and F1 score of 0.903 +/- 0.01. The generalisability error was computed as the difference between the internal and external model performances obtained for each test set.

The test set of lowest chemical diversity was built to mimic a standard model training and evaluation scenario, where training and test data are generated from the random split of an initial dataset. Accordingly, both of these sets were randomly drawn from the same initial pool of molecules (all those up to 200 grams per mol) with no replacement, and therefore are expected to be characterised by the same data distribution. This means that the resulting model was evaluated with molecules whose characteristics are similar to those used in the training process. Consequently, the model was able to generalise to the test molecules with a performance on par with the levels of its internal assessment.

On the other hand, the remaining test sets were built to mimic deployment scenarios. Although real-world sets can include virtually anything, in this controlled experiment test sets were designed to be increasingly different from the training set. Disregarding the characteristics of the model input led to the increase of the generalisability error of the model, which grows with the extent of the stationarity assumption violation. In our experiment, this went as far as a drop of 11% between internal and external model performance.

In conclusion, our findings suggest that models do not generalise indefinitely across chemical space. Models are produced from finite collections of examples and their predictive capability often relates to the degree of conformity of inputs to the training data. Therefore, test [external] performances should only be expected to be on par with internal evaluation results when test inputs comply with the stationarity assumption. Ultimately, this exercise illustrates why deployed models may not perform up to the levels reported during their publications, and motivates the need for the identification of the applicability domain.

4.2.2 Generalisability Tool

4.2.2.1 Generalisability tool description

In this section, we address the generalisability of a given QSPR model. The concept of generalisability refers to the extent to which models are able to generalise to a new input, and most related approaches aim to identify and delimit the regions of chemical space where models hold their predictive power. Here we revisit the notions of training data distribution, stationarity assumption, and anomaly detection to develop the generalisability tool (GT), designed to assess whether a new input example is likely to belong to the applicability domain of a given model, based on its univariate and multivariate anomaly character.

The tool implementation requires restructuring the previous outlier detection pipeline. Under this new strategy, domain outliers will still be part of the data processing step. That is because their filtering process is dependent on the feature content of each dataset and requires the specification of the upper and lower bounds of each feature domain. On the other hand, univariate outlier treatment is now embedded in the generalisability tool as the first detection step, and carried out using the same method and threshold as before. The second step of the tool is dedicated to the identification of multivariate outliers. A simplified representation of the generalisability tool is shown in **Figure 11**.

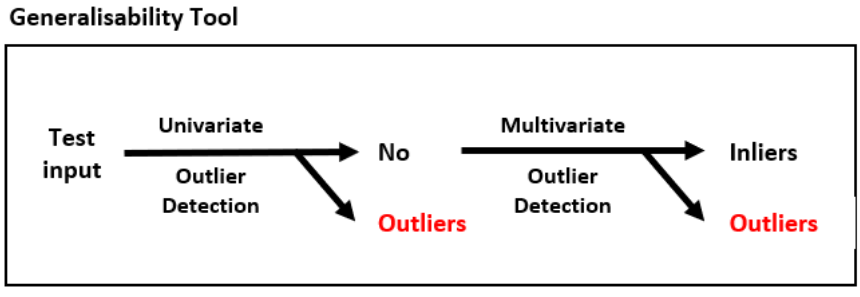


Figure 11 The generalisability tool.

Multivariate outlier detection methods adopt different strategies, whose success seems to be context-dependent.[213,230–232] Contrary to other AD solutions that are bound to specific outlier detection approaches, the generalisability tool is a general-purpose solution which through the consideration of different candidate multivariate outlier detection alternatives, is able to easily identify and implement the one that is most appropriate for a particular QSPR problem.

Generalisability studies must be carried out before using the GT for the assessment of any new unlabelled examples. This step aims to assess whether a model can benefit from this tool, *i.e.* whether any of the *in-built* multivariate outlier detection methods can correctly identify any incoming outliers with respect to the QSPR problem at hand. These studies also aim to identify the range of anomaly thresholds for the reliable use of the tool, and help select the one that is most suitable for a particular GT application.

Generalisability studies are based on the original QSPR model, training and test sets. Both training and test data are expected to be cleaned, scaled, and treated for domain and extreme outliers. The latter ensures that the success of each multivariate outlier model is not influenced by extreme outliers, as these would not be present by the time data is passed onto the second stage of the generalisability tool. Additionally, the training set should also be class balanced with respect to the target variable.

The training data are used for internal performance assessment, multivariate outlier detection model fitting, and anomaly threshold estimation. Internal performances are measured as the mean F1 score of a 3-fold cross-validation routine on the training set, using the original QSPR model estimator and hyperparameters. These act as the reference (or best-case) level which the model may be expected to operate at within its applicability domain. Then, six *in-built* multivariate outlier detection methods are fitted to the training set - Isolation Forest[238,239], Local Outlier Factor[240], One class SVM[241], Angle-Based[242], Histogram-

based[243], and Mahalanobis-Based[244] outlier detection methods. A summarised description of each estimator and their training considerations can be found in ***Supplementary Information H***. After fitting, each outlier detection model is used to evaluate the anomaly character of all training examples. Anomaly thresholds are derived from the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} quantiles of the distribution of anomaly scores of the training data. In theory, these provide several anomaly threshold references that test examples should comply with in order to be considered part of the model's applicability domain.

On the other hand, the entire test set is also processed by each outlier detection model and then randomly sampled five times (with replacement). The sampling strategy adds robustness to the study and sample sizes are set to ensure that meaningful performances can be calculated from each sample. All test samples are trimmed according to each threshold value, balanced, and evaluated by the original QSPR model. An unbounded scenario is also considered, where no threshold is implemented. The mean F1 score obtained across the five test samples is held as the model performance enabled, when a given anomaly threshold is adopted. This procedure is repeated for all multivariate outlier detection models and the idea is to identify those whose anomaly scores correlate well with the performance of the QSPR model.

The results are presented in a table, where each row corresponds to the results for a single outlier detection model and the columns (left to right) contain the QSPR model performances on test sets of increasing anomaly content. Suitable models are those that show an increase in test performance as the threshold becomes stricter [smaller], meaning that the outlier estimator can successfully learn the training data distribution and identify outliers that do not comply with it. Note that sometimes, the lowest thresholds do not enable enough examples in the output from which meaningful performances can be derived. In these cases, the trend assessment should not consider such threshold performance results.

The best model is the one amongst those that enable the highest test performance for stricter [smaller] thresholds, such that the applicability domain is maximised.

After manually selecting the multivariate outlier detection framework, the generalisability tool is ready to be used. If no suitable method can be identified, the generalisability tool is declared unfit for the QSPR prediction task. Note that the list of multivariate outlier detection estimators can be easily extended with any implementations that are compatible with our Python scripts. A schematic representation of the generalisability tool within the QSPR prediction pipeline is shown in **Figure 12**.

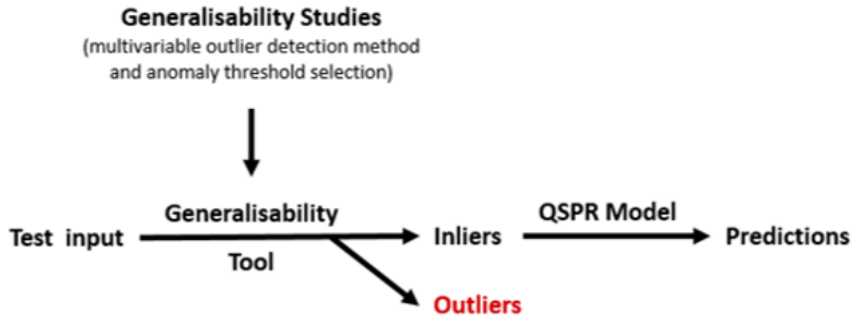


Figure 12 The generalisability tool within the QSPR prediction pipeline.

4.2.2.2 Generalisability tool implementation

In this section we provide a step by step implementation of the generalisability tool. Here, we recreate the controlled deployment scenario described and used in ***Chapter 4 / 4.2.1 Chemical diversity studies*** under the same assumptions, to illustrate the impact of the GT on the output performance when a deployed model faces a series of unseen inputs, which are not representative of its training data. Refer to ***Chapter 4 / Methods / Crystallizability Dataset & Model*** for details on the deployment scenario implementation. Once again, note

that only the crystallizability dataset is large enough to be used at this stage. Accordingly, ten subsets were considered: one training set of molecules up to 200 grams per mol and nine test sets whose molecular weight content is increasingly different from the training data. Note that the lowest chemical diversity test set was randomly sampled from the same pool used to build the training set and so both are expected to follow the same data distribution.

First, generalisability studies were carried out using the training set and the test set of lower chemical diversity in order to assess whether the crystallizability model can benefit from this tool. These studies were carried out as described in **Chapter 4 / 4.2.2.1 Generalisability tool description.** The internal crystallizability model performance enabled a F1 score of 90 +/- 1% and the generalisability study results are shown in **Table 8**. All models were able to achieve or surpass the level of internal model performance when assessed (**Table 8**, at 90^{th} quantile and Unbounded). However, the *hbos* and *ocsvm* models do not seem to produce anomaly values that correlate well with model performance, making them unfit for this QSPR problem. All other models achieve similar highest performance levels (**Table 8**, at 10^{th} quantile) and comply with the expected trend, but the *abod* model shows the steadiest one of them all. Hence, the *abod* model is selected as the most suitable multivariate outlier detection method and the generalisability tool assembly is completed.

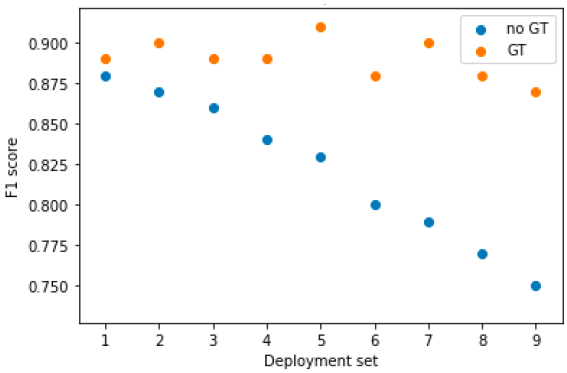
Table 8 Generalisability study results for the crystallizability case study. The multivariate outlier detection (MOD) model key is as follows: Isolation Forest[238,239] (if), Local Outlier Factor[240] (lof), One class SVM[241] (ocsvm), Angle-Based outlier detection[242] (abod), Histogram-based outlier score[243] (hbos), and Mahalanobis-Based outlier detection[244] (mahalanobis). See ***Supplementary Information H*** for details on the estimators. Each column corresponds to the mean F1 score performance on five random sub-samples of the original test set, after being trimmed for different anomaly threshold values. Anomaly thresholds were derived from the collection of anomaly scores computed for all training examples. These correspond to the 10th, 25th, 50th, 75th, 90^{th,} and 100th (unbounded scenario) percentile of the anomaly score distribution. See ***Chapter 4 / 4.2.2.1 Generalisability tool description*** for details on how to conduct the generalisability studies.

MOD Model	10th qt	25th qt	50th qt	75th qt	90th qt	Unbounded
abod	0.95	0.93	0.92	0.91	0.91	0.91
mahalanobis	0.96	0.93	0.91	0.91	0.91	0.91
hbos	0.90	0.88	0.89	0.89	0.90	0.91
if	0.95	0.93	0.91	0.90	0.91	0.91
lof	0.94	0.92	0.91	0.91	0.90	0.91
ocsvm	0.89	0.90	0.90	0.90	0.91	0.91

Then, we evaluated the model performance on each test set, before and after processing it with the generalisability tool. Here, the first assessment mimics a deployment scenario where no considerations are made about the input to be fed to the QSPR model, whilst the second mimics the scenario where inputs are narrowed down to the model's applicability domain. Results are shown in **Figure 13**, which includes a subfigure A and a table B. As expected, the increase in test set content that does not belong to the model's applicability domain is directly proportional to the decrease in test model performance (**Figure 13A**, blue dots). However, after treating each test set with the GT, *i.e.* filtering it for the molecules that the model should be able to generalise to, test performance is restored to the internal assessment levels every time (**Figure 13A**, orange dots). Furthermore, the fraction of dataset recovered roughly corresponds to the fraction of AD molecules initially included in each test set (**Figure 13B**). Although these results reassure us about the performance of the GT on the crystallizability prediction problem, note that in real world

scenarios the threshold value does not reflect the fraction of input examples returned by the generalisability tool.

A



B

Deployment set	% mol content (mw < 200g/mol)	Input % in the output
1	90	83
2	80	74
3	70	66
4	60	57
5	50	49
6	40	40
7	30	32
8	20	24
9	10	15

Figure 13 Deployment scenario results on the crystallizability dataset. **A** Each orange/blue pair corresponds to the F1 score performance obtained for a different ''deployment'' input set, before and after being processed by the generalisability tool. Deployment sets 1 to 9 correspond to test sets of 5000 examples, out of which 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, and 10% of their content corresponds to molecules with a molecular weight up to 200 grams per mol. Refer to ***Chapter 4 / Methods / Crystallizability Dataset & Model*** for details. The standard deviation of 0.01 for each point is not represented in the plot. **B** Lists the percentage of molecules up to 200 grams per mol in each dataset, and the percentage of examples recovered after filtering down initial inputs using the generalisability tool with an anomaly value threshold equivalent to the 90th quantile of the training anomaly values distribution.

During model development, the amount of labelled data available is often limited and although experiments like the one above may not be feasible, generalisability studies can still be carried out with the same intent as before: assess whether the given QSPR model could benefit from the generalisability tool, identify the most suitable multivariate outlier method for the task, and assist in the selection of a suitable generalisability threshold. This analysis was conducted for the HBND, crystal habit and polymorphism datasets. The corresponding datasets used in this experiment were those produced and/or described in ***Chapter 2 / 2.2.3 The models.*** Generalisability study results can be found in ***Supplementary Information*** I. Generalisability studies.

Threshold values are often selected to process inputs with the same predictive power exhibited during the internal model assessment. For models with high predictive power, we suggest that the 90th quantile of the distribution of training anomaly values is a good default threshold. However, limited performing models were found to require stricter thresholds to operate at their expected accuracy levels, due to potential data-inherent faults in the outlier detection model. In these cases, good threshold alternatives can be derived from the corresponding generalisability studies.

Usually, as the generalisability threshold decreases, the model performance on the output increases (**Table 8**). That said, thresholds must be selected on a QSPR project basis and tuned to the requirements and specifications of each prediction task. For example, virtual screening algorithms can afford to adopt lower threshold values to produce a large pool of potential hits for further analysis, whilst stricter boundaries must be used for individual molecular property prediction or characterisation purposes.

Finally, we note that multivariate outlier detection-based approaches should ideally be carried out after feature selection, since the performance of most detection methods tend to deteriorate in high dimensions. Furthermore, solely focusing on the features that are critical to the prediction task prevents us from discarding points whose anomalous character is based on other less relevant features.

In conclusion, the results of this study suggest that out generalisability tool is able to restrict QSPR models to their corresponding applicability domain, regardless of the initial collection of molecular inputs at hand.

4.2.3 Confidence Tool

The material in this section has been published in *Frade et al.*[104] and the text was adapted to the fit with the flow of the chapter.

4.2.3.1 Confidence tool description

In this section, we address the confidence aspect of the applicability domain of a given QSPR model. The concept of prediction confidence refers to the level of trust with which the model computes individual predictions, and most related approaches aim to identify and delimit the regions of chemical space where models output confident results. Here we revisit the notions of probability of class membership and confidence estimation to develop the confidence tool, designed to evaluate the confidence of individual predictions from calibrated probabilities of class membership.

Some models naturally produce the likelihood of the predicted outcomes as part of their results, whilst others provide a certainty score that can be recalibrated into a probabilistic scale. For the latter, class probabilities are often obtained from models trained to map native certainty scores to probabilities in the interval [0, 1]. Moreover, these probabilities must be calibrated to reflect the true likelihood of events. For binary problems, calibrated probabilistic outcomes are often achieved by fitting a sigmoid[245] or an isotonic[246] regressor to a class balanced version of the training set under a cross validation routine.[247] The multi-class case adopts the extended versions of these methods[248] based on the 'one vs rest' classification fashion.[249]

Our goal was to build a tool extended to multi class problems that provides functionality beyond simple confidence estimation. In these settings, each prediction output consists of the array of calibrated class membership probability estimates. A confident threshold corresponds to the minimal gap that must exist between the two highest probabilities in this array, and a confident guess occurs when this gap is satisfied.

The algorithm behind the confidence tool takes in a predictive model, some test data, and a given prediction confidence threshold as input. The model must be a classifier able to output calibrated class probabilities in terms of the true posterior probability. The confidence

threshold must be a number in the interval [0, 1], where 0 means that no confidence is considered during the prediction stage, and 1 means that the model must assign the example to a category only when its class probability is 100%.

A simplified representation of the confidence tool is shown in **Figure 14**. Under the confidence tool framework, each test example is evaluated by the predictive model and the array of probabilities is obtained.

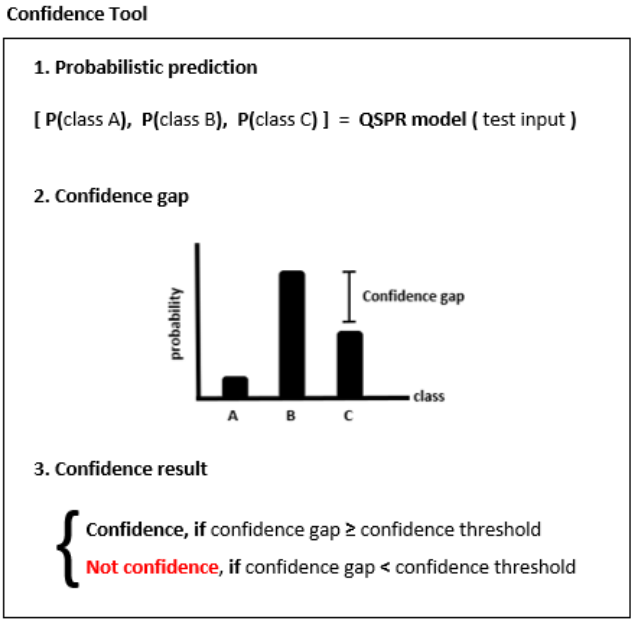


Figure 14 The confidence tool. The confidence tool obtains the probability of class membership for each new input. Probabilistic results are used to compute the prediction's confidence gap. This is then assessed against the confidence threshold to conclude whether the prediction was considered confident or not.

The gap between the two highest class probabilities of the output array is calculated and its comparison against the confidence threshold determines whether the prediction was confident. If so, the example is assigned to the most probable class. The test examples that the estimator is not able to predict with confidence are not classified. These are stored and

can be passed onto another model, or to the same model exposed to a lower confidence threshold. A schematic representation of the confidence tool within the QSPR prediction pipeline is shown in **Figure 15**.

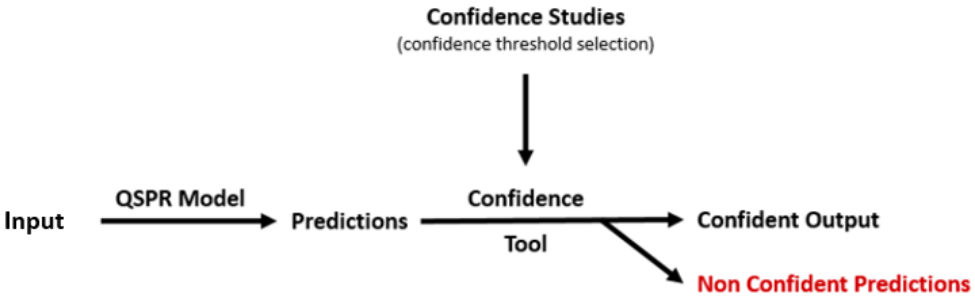


Figure 15 A schematic representation of the confidence tool within the QSPR prediction pipeline.

4.2.3.2 Confidence tool implementation

In this section, we use the HBND case study to illustrate the extended applicability of the confidence tool, as described in the original work.[104] The same analysis was conducted for the crystallizability, crystal habit and polymorphism datasets. Results can be found in ***Supplementary Information J***. The test sets and models used throughout this experiment were those produced and/or described in ***Chapter 2 / 2.2.3 The models*** and first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model***. Note that all properties are modelled using an SVM estimator, which is not a probabilistic framework. Fortunately, the third-party SVM estimator used for model learning can be set to automatically produce calibrated probabilities of class membership, which is internally achieved through Platt scaling[245] for binary problems or its multi-class extension[248], and the adoption of a class balanced training set.

First, we investigated the benefits of considering confidence thresholds at all. To determine how many predictions were facing a small probability difference between their two most probable classes, the test set was first evaluated by the model with no confidence threshold and then subject to a 5% confidence threshold. We found that whilst all HBND examples would be predicted under no confidence threshold, only 83% of the test set could be confidently predicted when the 5% confidence threshold was applied. This means that 638 examples were being assigned to a given class with no more than a 5% difference between the top two probability estimates. Running models with very low confidence thresholds suggested that some of the correct answers that the model outputs when no thresholds are implemented turn out to be lucky guesses. This sensitivity implies that the performance of models with no minimal confidence restriction can rapidly decrease when faced with noisier datasets. Thus, we conclude that confidence threshold implementation is an efficient way to improve the robustness and reliability of a model.

Then, we carried out confidence studies to test the effect of increasing confidence thresholds on the fraction of test examples that a model can predict with confidence and the corresponding accuracy. The model was used to predict HBND labels for the complete test set under different confidence thresholds. The results are shown in **Figure 16**.

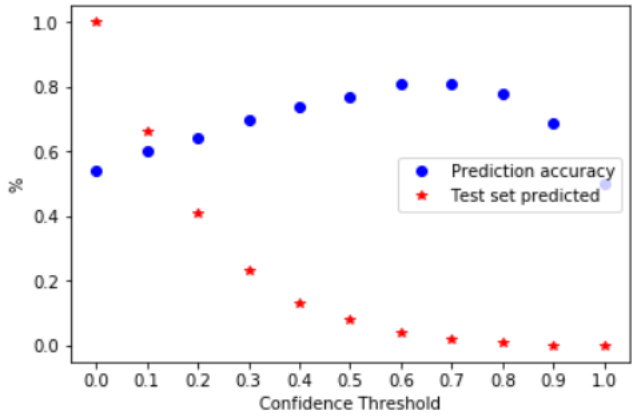


Figure 16 Confidence study results for the HBND dataset. The plot shows the effect of confidence thresholds on the percentage of the HBND test set predicted with confidence and corresponding prediction accuracy. The internal assessment of the HBND has an accuracy and F1 score of 0.55 +/- 0.01.

For each confidence threshold used, there is a pair of red and blue dots representing the percentage of test examples that were predicted with confidence and corresponding prediction accuracy. Generally, as the confidence threshold increases, the accuracy of confident estimations also increases and does improve considerably. Conversely, the percentage of test examples that the model can predict with confidence drops rapidly. For example, whilst the absence of a confidence restriction allowed the model to predict the complete test set with an accuracy of 55%, a 0.3 threshold enabled the model to output predictions for 23% of the test set with an accuracy of 70%, or for 4% of the test set with an accuracy of 81% when the threshold was increased to 0.6. Results obtained for the other case studies show that these trends are consistent (***Supplementary Information J***).

When the number of confident predictions gets too small, meaningful statistics about the general performance of the model cannot be derived. As shown in **Figure 16**, the prediction accuracy for confidence values above 0.6 is obtained from a small collection of examples that may no longer be a good representation of the original data distribution. In

the HBND case, 0.6 is the maximum confidence threshold to be adopted for the computation of a meaningful overall model performance (which considers at least 100 test examples). We stress that despite this apparent upper bound, the confidence associated with the outputs obtained at high thresholds is still valid. In summary, confidence thresholds make it possible to operate the model up to any desired level of accuracy within the bounds discussed, however, a compromise between accuracy and access to answers is required.

Finally, we use the confidence tool to predict the test set over seven classification rounds of decreasing confidence thresholds. The idea was to feed into a given classification round all the test examples that the model was not able to confidently predict in the previous round. We also note that the confidence threshold can be continuously decreased, as long as each round of classification still performs better than random. The results are shown in **Table 9**.

Table 9 Results per round of the SVM model for HBND prediction, subject to a sequence of seven confidence thresholds. In this setup, test examples were being predicted, and only the ones that could not be predicted with confidence during one round would be passed onto the next (of lower threshold) for evaluation. Accuracies are calculated per round. The internal assessment of the HBND has an accuracy of 55 +/- 1% and a random baseline performance of 25% on a balanced test set.

Round	Confidence Threshold	Confident Predictions	Correct Predictions	Accuracy (%)
1	0.6	138	112	81
2	0.5	145	107	74
3	0.4	197	134	68
4	0.3	370	242	65
5	0.2	634	355	56
6	0.1	933	500	54
7	0.0	1235	538	44

As expected, gradually relaxing the confidence restriction enables the estimation of progressively less confident new answers at each round, which increases the fraction of the

test set predicted. The true value of this approach is its ability to accommodate any number of rounds and threshold values, such that the number of predictions can be maximised whilst controlling the overall performance. Moreover, the confidence associated with the output answers of a given round is known to lie between the confidence threshold that the current and previous rounds operate under. Thus, the possibility of fine-tuning the confidence threshold step between rounds allows the discrimination of predictions based on their level of prediction trust at any desired resolution. Lastly, this setup can be extended to the prediction of test data using different models at each round, subject to the same or different confidence thresholds. The idea is that different classifiers may identify different patterns in data according to which they predict new examples. Processing test data this way was found to successfully maximise the number of confident predictions from the HBND test set for a given confidence threshold (results not shown).

In conclusion, we believe that the confidence tool offers the possibility of tailoring the performance of a given probability-generating classification model to the risk and cost requirements of each project, but a compromise between model performance and access to answers is required. As a final remark, we highlight the fact that confident predictions do not always lead to correct answers.

4.2.4 Applicability domain

In this final section, we bring the generalisability and confidence tools together in order to test them as a unified pipeline for the identification of the applicability domain of models of high and limited predictive power. Here, we assess model performance on each case study – crystallizability, HBND, crystal habit, and polymorphism – under four different settings: no consideration of the model input, under the generalisability tool alone, under the confidence tool alone, and under both tools.

The crystallizability dataset and model were prepared as described in the first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model***. The remaining datasets and models used throughout this experiment were those produced and described in ***Chapter 2 / 2.2.3 The models.*** Note that the train and test objects of each individual case study were obtained using the same strategy (refer to ***Chapter 2 / Methods / Data Collection*** for details). Thus, all input objects comply with the intended use of each model, which verifies the first dimension of the applicability domain. Furthermore, training and test sets were randomly sampled from the same initial pool of molecules, and so both are expected to follow the same data distribution.

Generalisability and confidence studies were carried out for all cases in earlier sections, and corresponding results can be found in ***Supplementary Information I*** *&* ***J***. From these, the generalisability and confidence threshold selection aimed to identify the optimal applicability domain at each case. **Table 10** summarises the settings of both tools adopted for each case study. Note that the anomaly thresholds of HBND, crystal habit, and polymorphism are based on stricter [lower] quantiles, due to the limitations of the predictive power of their models.

Table 10 Generalisability and confidence tool settings adopted for different case studies. Generalisability tool settings were selected based on the results and suggestions described throughout ***Chapter 4 / 4.2.2.2 Generalisability tool implementation*** and ***Supplementary Information I***. Confidence tool settings were selected based on the results and suggestions described throughout ***Chapter 4 / 4.2.3.2 Confidence tool implementation*** and ***Supplementary Information J***. The outlier detection method key is Angle-Based outlier detection[242] (*abod*), One class SVM (*ocsvm*), and Isolation Forest (*if*). The anomaly threshold corresponds to the quantile of training anomaly score distributions. Refer to the mentioned sections for more details.

Case study	Generalisability Tool		Confidence tool
	Outlier detection method	Anomaly threshold	Confidence threshold
Crystallizability	*abod*	90th quantile	0.5
HBND	*abod*	50th quantile	0.3
Crystal Habit	*ocsvm*	50th quantile	0.2
Polymorphism	*if*	75th quantile	0.2

The applicability domain results are shown in **Table 11**.

Table 11 Applicability domain studies. The internal performance results were borrowed from ***Chapter 2 / 2.2.3 The models*** section.

		Crystallizability	HBND	Crystal Habit	Polymorphism
Internal performance (mean F1 score)		0.95 +/- 0.02	0.55 +/- 0.01	0.49 +/- 0.02	0.62 +/- 0.02
No method	F1 score	0.93	0.55	0.49	0.62
	Input (%)	100	100	100	100
GT	F1 score	0.95	0.56	0.50	0.65
	Input (%)	91	51	50	74
CT	F1 score	0.97	0.57	0.59	0.67
	Input (%)	90	20	32	42
GT & CT	F1 score	0.97	0.65	0.60	0.69
	Input (%)	82	6	20	37

The ‘No method’ scenario led to performances similar to those achieved during internal model assessment, because each training and test data pair was produced by the random sampling of the same pool of molecules and therefore must follow a similar distribution (refer to ***Chapter 4 / 4.1.2 The stationarity assumption of statistical learning*** for details). Moreover, the individual GT enabled the slight but consistent improvement in model performances across all case studies. Again, models were expected to generalise well to most of their corresponding test examples for the same reason just mentioned. The content similarity of each training and test set pair can be further confirmed when comparing the fraction of input retrieved by the GT to the training anomaly quantile employed as the generalisability threshold. On the other hand, the CT was found to enable useful increases in performance across most cases. Unsurprisingly, since confidence estimation uses the trained classifier, the CT is expected to work more efficiently with respect to model error reduction. Finally, the best results were obtained when both tools were used in tandem. In fact, restricting QSPR models to reliable inputs and confident predictions boosted model performances to levels beyond their internal benchmarks, with the associated cost of fewer predictions.

In summary, these findings suggest the practicality and performance of both GT and CT tools, as well as the importance of considering all aspects of a model’s AD during any prediction task.

4.3 Conclusion

The chemical space is amazingly large when compared to the set of molecules used to train an estimator. Chemical datasets are intrinsically biased towards their molecular content, and consequent models should not be expected to generalise indefinitely across chemical space. The fundamental assumption at play states that training, test and real-world

objects must all be drawn from the same data distribution. Chemical diversity results were carried out to show how breaking this assumption leads to the loss of model generalisability and consequent decrease in output performance, by testing models against molecules drawn from distributions increasingly different from the training set.

Setting the basis for why most QSPR models seem to underperform after their deployment, motivates the need for the identification of applicability domains.[60–63] Applicability domain strategies address the validity of inputs, generalisation capacity of models, and confidence of predictions by ensuring that unseen objects comply with the intended use of the model, follow the training set distribution and provide enough evidence for its confident classification. This young and active area of research still lacks general, robust, and ready to use solutions. Thus, we provide our contribution to the field by reporting the generalisability and the confidence tool. This combination of measures was developed as a general-purpose and easily implementable strategy to describe the applicability domain of binary and multi-class classifiers.

The generalisability tool identifies the regions of chemical space where a given model holds its predictive power by assessing the extent to which it is able to generalise to each new input object. This consists of comparing each test molecule to the training set characteristics, determine its anomaly character and compare it against a generalisability threshold. This approach was designed to guarantee that chemoinformatics classifiers are used under appropriate contexts. Generally, as the generalisability threshold decreases, the accuracy of reliable estimations increases, but the percentage of test examples that the model is able to predict reliably drops rapidly. Contrary to other implementations,[224–226] our GT is a general-purpose solution (*i.e.* it is not bound to a single outlier detection method) whose applicability and performance seems to be less dependent on the QSPR problem at hand.

The confidence tool identifies the regions of chemical space where a given model produces confident results, by assessing the level of trust with which the model outputs individual predictions. This consists of estimating the confidence of each prediction based on class probability memberships and compare it against a confidence threshold. Generally, as the confidence threshold increases, the accuracy of confident estimations also increases, but the percentage of test examples that the model is able to predict with confidence drops rapidly. The expanded functionality of our CT allows you to estimate and apply prediction confidence results in a number of useful ways.

Tool validation was carried out for each case study. The tools were found to work best when used in tandem and the corresponding output predictions were able to achieve or surpass the level of internal model performance in all cases.

The proposed strategies demonstrate that controlling the quality of inputs and outputs can help to identify the regions of chemical space where QSPR models hold their predictive power and produce confident predictions. Again, most chemoinformatics models may not perform up to industrial standards, yet we believe that those performing reasonably better than random have an underestimated potential that is often left unexplored. Furthermore, we also believe that under these constraints, models of limited performance are able to add great practical value.

4.4 Methods

Crystallizability Dataset & Model

An extended version of the crystallizability dataset was acquired from our CCDC collaborators, which retrieved molecules and labels using the same principles described in the original article.[32] High-level molecular representations (refer to ***Chapter 2 / 2.2.2 The datasets*** for details) were calculated for all 440,653 molecules and the final dataset was

cleaned, scaled and treated for domain and extreme outliers as described in ***Chapter 2 / Methods / Data Processing***. The modelling framework (base estimator and hyperparameters) was optimised following the same approach as described in ***Chapter 2 / Methods / Machine Learning Models***. This corresponded to an SVM with and RBF kernel, with a C parameter of 10 and a gamma parameter of 0.001, with a reported internal assessment of 0.95 +/- 0.02 accuracy and F1 score. At times, the entire crystallizability dataset was represented by random sample of 20,000 molecules to accelerate the experiments.

Controlled deployment scenario. In some experiments is useful to break down a large dataset into smaller subsets with specific characteristics. This controlled set up is designed to mimic a deployment scenario, where models are exposed to several different inputs. Molecular weights were recalculated for all entries using RDKit methods, to be used as a measure of chemical diversity. Training and test sets were produced through the random sampling of the processed dataset, under different molecular weight constraints. Accordingly, the entire crystallizability dataset was separated into two subsets: one with all molecules with a molecular weight up to 200 grams per mol and another with all remaining molecules. The training set comprised a collection of 10,000 examples randomly sampled from the first subset, from which all training examples were immediately removed to avoid cross contamination between training and test data. Eleven test sets of 5000 examples were produced, out of which 100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, 10%, and 0% of their content corresponded to molecules sampled from the first pool (subset of molecules with molecular weight up to 200 grams per mol) and the remainder was sampled from the second. Refer to the **Table 12** below for clarification. All sampling steps were at random. Finally, training data was balanced with respect to the target variable.

Table 12 Details of the test data used under the controlled deployment scenario. The table clarifies the composition of each test set. The training set is produced beforehand and it only contains molecules of pool 1. The sampling of each pool occurs at random and with replacement, after discarding all training examples from pool 1.

Test set	% mol content (Pool 1: mw < 200g/mol)	% mol content (Pool 2: mw > 200g/mol)
1	100	0
2	90	10
3	80	20
4	70	30
5	60	40
6	50	50
7	40	60
8	30	70
9	20	80
10	10	90
11	0	100

Chemical Diversity Studies

Training and test data were obtained as described in ***Chapter 4 / Methods / Crystallizability Dataset & Model***. The training set was used for model learning and internal model assessment. The internal model performance was measured as the mean accuracy and F1 score of a 3-fold cross-validation routine.

External model performances were obtained by evaluating the predictive model on each test set. The generalisability error was computed as the difference between the internal and each of the external model performances. Generalisability error values were plotted against the average molecular weight of each corresponding test set.

Chapter 5 – Deriving chemical insights

In Chapter 5, we attempt to derive chemical insights about structure-property relationships from QSPR models. Here, we cover concepts such as feature selection and model interpretability. The key outcomes include a feature selection strategy that identifies the structural aspects that most drive the prediction of a particular property, and a model interpretation pipeline that reveals how changes in the molecular structure may affect the predicted property value. Our studies demonstrate that given the right tools, it becomes possible to overcome the curse of dimensionality, whilst a new way to advance our chemical understanding is also unlocked. Unfortunately, the latter continues to be restricted to models with high predictive power.

5.1 Introduction

5.1.1 The curse of dimensionality

In QSPR studies, inputs (descriptors) attempt to describe the aspects of molecular structures that seem to be associated with a given property or behaviour. The relevance of the structural description determines the success of the mapping task, but structure-property relationships are often unclear and the selection of pertinent features is not trivial. Molecular descriptors are numerous and specific. In theory, more features should enable a more complete description of the overall molecule and consequently increase the likelihood of covering the relevant structural information for the exercise. However, considering all options at once imposes technical issues that may compromise the quality or even the feasibility of the QSPR modelling task.[80] In fact, high dimensionality is a recurrent problem in QSPR datasets. The volume of feature space grows exponentially with the number of features

and data becomes increasingly sparse, causing most machine learning methods to break, statistical concepts and metrics to become less useful, and results to become less significant.[250] This problem is often referred to as the *curse of dimensionality*, whose consequences include the general increase of noise in data, multicollinearity between features, computational cost of the mapping task, and chance of model overfitting.[251]

Many molecular descriptors are also derived from one another (*eg. PEOE_VSA, SMR_VSA, SlogP_VSA, or Estate_VSA* are all derived from the van der Waals surface area and some other descriptor) and/or highly correlated (*eg.* molecular weight and heavy atom, bond, or functional group count). Conveniently, amongst redundant and irrelevant features, only a manageable fraction of those initially considered tend to be useful for the effective prediction of the target variable.

5.1.2 The feature selection process

Feature selection (FS) refers to the process of identifying the subset of features that are useful for a given machine learning problem. Feature selection techniques efficiently reduce the dimensionality of real-world datasets, acting as powerful solutions against the curse of dimensionality and its consequent drawbacks. Besides, the procedure has also shown to facilitate the visualisation and understanding of data, enable a more accurate definition of the model's applicability domain, enhance model generalisability, and improve model interpretability.[252–254]

There are three main approaches to supervised feature selection: embeddings, filters, and wrappers.[250] In short, embeddings intrinsically select features as part of the learning procedure. Filter methods are implemented before model training to assess each feature individually and rank them according to some measure of importance. Wrappers also operate during the model training process. Unlike embedded or filter methods, these are

usually not bound to a particular model estimator. Furthermore, they also tend to consider both feature and model hypothesis spaces during the feature selection process, as well as the interdependencies between input features. Numerous reviews and surveys on feature selection approaches and applications can be found across the literature.[156,158,220,255–262]

Feature selection must not be confused with other dimensionality reduction techniques based on the projection or compression of features. Feature selection preserves the original feature space and simply selects the subset of features found to be most useful for the machine learning task. Here, we revisit the notions of redundancy and relevance under the context of feature selection.[80,263] Accordingly, feature selection techniques are adopted under the assumption that data may contain redundant and/or irrelevant features, which can be removed with no information loss to the system.[256]

Feature redundancy represents the frequency with which the information that a feature adds to the dataset reappears, ranging from unique to ubiquitous.[156] In this context, features that hold similar information are deemed redundant with respect to each other. Feature redundancy-based approaches focus on minimising the redundancy contained in the dataset, generally by identifying groups of redundant features and narrowing them down to single representatives.[80] These methods are often univariate and assess redundancy based on some measure of feature correlation or distribution similarity.[80,263] Popular metrics include Pearson and Spearman rho correlations, Shannon entropy, and mutual information.[259] However, methods often do not consider the interaction between features and ignore the fact that correlation does not imply redundancy, as demonstrated by *Guyon et al.*[259]

On the other hand, feature relevance-based approaches are focused on identifying the most relevant features for a given prediction task.[80] In this context, features range from critical to irrelevant based on the extent of their contribution to the discrimination of model

outputs.[156] Relevance is often assessed by the decrease in model performance when the feature is discarded from the modelling task.[158] Wrappers tend to offer the most flexible and accurate feature relevance assessment.[220] Such techniques often explore the feature space by assessing subsets of features via cross validation under the original modelling framework.

Different types of approaches are often combined to enable feature selection within a minimum-redundancy maximum-relevance framework.[80,264] Understanding the level of relevance and redundancy character of each feature in a given QSPR modelling task is an important step towards avoiding the curse of dimensionality and rationalising the model's prediction mechanism.

5.1.3 Interpreting QSPR models

Beyond the ability to predict properties from molecular structures, it is of interest to understand how changes to the structures may affect the property value. Model interpretation refers to the clarification of the prediction process through which models transform and combine input features to estimate output labels. The analysis not only increases model transparency and user trust, but it also generates insights that may contribute towards the understanding of structure-property relationships.

It is important to note the difference between the input-output variable mapping and the real nature of the underlying systems. Models rely on patterns found in data, *i.e.* useful correlations that the estimator was able to identify amongst the input features provided. For that reason, the mechanism by which structural descriptors are processed to generate predictions does not necessarily reflect the reasons and means by which molecular structures adopt certain properties and behaviours. Simply put, correlation does not imply causation. Nevertheless, the insights derived from data driven models are still extremely valuable and often adopted as guides to action. In QSPR studies, these findings reveal how

different structural features influence the prediction of molecular properties, acting as useful information for model validation by domain experts, guides for the screening of virtual libraries,[265,266] pointers for the design and optimisation of molecules,[267] or supporting claims for the proposal or disproof of new mechanistic hypothesis and theories.[4,123,268–272]

Most QSPR models are based on estimators that are not easily interpretable. Thus, the scientific community developed several solutions, which can be divided into model estimator dependent or agnostic. A survey on these can be found in *Polishchuk et al*.[57] Estimator dependent solutions comprise knowledge extraction strategies particularly designed for specific estimators, such as multivariate linear regression,[273] random forests[274] or neural networks.[275,276] On the other hand, model agnostic approaches do not rely on the internal structure of models. Instead, these treat models as black boxes and simply investigate how the output is influenced by changes in the input. Such investigations consider variable importance, sensitivity analysis, and partial derivatives. Variable importance estimation methods provide insight on the magnitude of influence that individual or groups of input features exert on the output.[277] Sensitivity analysis explores the nature of the relationship between the output variable and individual input features.[278] Partial derivatives measure the combined nonlinear influence of groups of inputs on the output.[279] SHAP[280] and LIME[281] are examples of popular, powerful and open source model-agnostic interpretation strategies.

The key to model interpretation lies within the quality of the model and data used to derive insights. Firstly, the trustworthiness of the knowledge extracted is determined by the accuracy of the mapping task, meaning that interpretation should be restricted to models with a reasonable predictive power.[63,69] Those of lower performances suggest that structure-property relationships could not be efficiently captured, which can occur for various reasons. Secondly, insights are derived from the investigation of the prediction process, which is

optimally enlightened by the examples that best follow the structure – property relationship captured by the model. Thus, model interpretation should also be restricted to the most relevant features driving the prediction task, as well as derived from and only applied to examples within the applicability domain of the model.[57] However, note that molecular features span across different levels of abstraction. Unfortunately, the development of QSPR models still tends to prioritise predictive power over interpretability, hence the success of the interpretation may also be limited by the level of abstraction of highly predictive features.[67,68]

5.1.4 Wrap up

In the last chapter we attempted to increase the value of our QSPR models by motivating the concept of applicability domain and proposing new strategies to restrict classifiers to regions of space where they can be used reliably. The adoption of these approaches led to the identification of contexts under which our limited performing models were able to add great practical value.

The ability to trust individual predictions encourages us to explore the prediction mechanism of models. In this final chapter, we attempt to derive chemical insights about structure-property relationships from chemoinformatics models. Here, we propose a novel approach to feature selection based on the minimum-redundancy maximum-relevance framework for the effective identification of the minimal set of most relevant features driving a given prediction task. We also implement a simple and model agnostic knowledge extraction pipeline designed to maximise model interpretability through the exploration of the magnitude and directionality of influence that input features exert on the target variable. The performance of both pipelines is investigated in light of each case study and the derived chemical insights are validated.

5.2 Results

5.2.1 Feature Selection

5.2.1.1 Feature selection pipeline description

The identification of structural features through logic and reasoning is a non-trivial task that, when feasible, requires substantial domain knowledge. This is because most structure-property relationships are unclear and structural features span across different levels of abstraction. Despite the many feature selection approaches reported in the literature, there is a lack of general-purpose implementations that select features within a minimum-redundancy maximum-relevance framework. Here we report a general-purpose feature selection pipeline designed to minimise the dimensionality of chemical datasets, returning the set of most relevant and minimally correlated features for a given QSPR modelling task. The system comprises a sequence of two *in house* developed feature selection methods aimed to address the relevance level and redundancy character of features in a dataset. A simplified representation of the feature selection pipeline can be found in **Figure 17**.

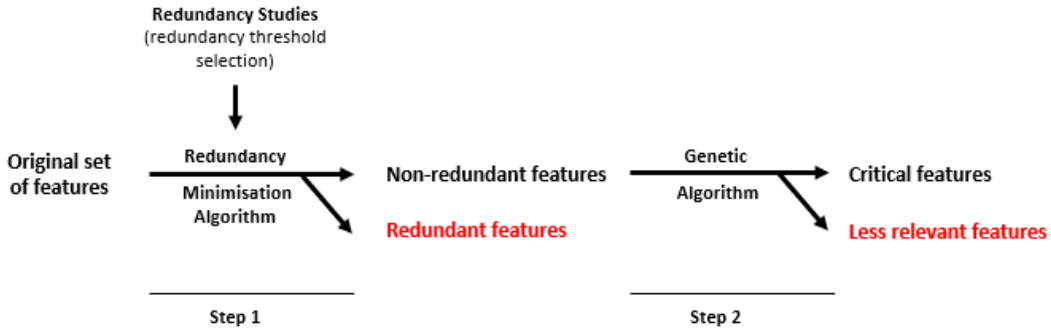


Figure 17 The feature selection pipeline. The redundancy and relevance character of features is assessed under a QSPR modelling context. The former (step 1) is addressed by the redundancy minimisation algorithm, whilst the latter (step 2) is taken care of by the genetic algorithm. At each step, the dimensionality (number of features) of the QSPR problem is expected to decrease with little to no loss of predictive power of the final model.

The first stage of the pipeline addresses feature redundancy. Identifying groups of redundant features allows us to narrow them down to single representatives with no information loss to the system. This exercise efficiently decreases the dimensionality of the problem, promotes the dataset compliance with the assumption of minimal feature multicollinearity adopted by most statistical methods, and allows us to reach consistent and invariable results during the identification of critical features. Our feature redundancy minimisation algorithm was inspired in the method proposed in *Peng et al.*[282] and is schematically represented in **Figure 18**.

Initially, explanatory features with zero variance are discarded from the dataset. These are features with a constant value across all the examples, adding no contribution to the distinction between target classes (**Figure 18, Step 1**). The remaining features are ranked based on their association strength with the target variable, and higher ranked features are assumed to have larger individual impact on the output values (**Figure 18, Step 2**). Following the order of the rank, the association of each feature with all of those below it is also computed. If the association strength rises above a predefined threshold, features are deemed redundant and the lower-ranked alternative is discarded and no longer considered (**Figure 18, Step 3**). Those that manage to hold their place in the rank form the solution set.

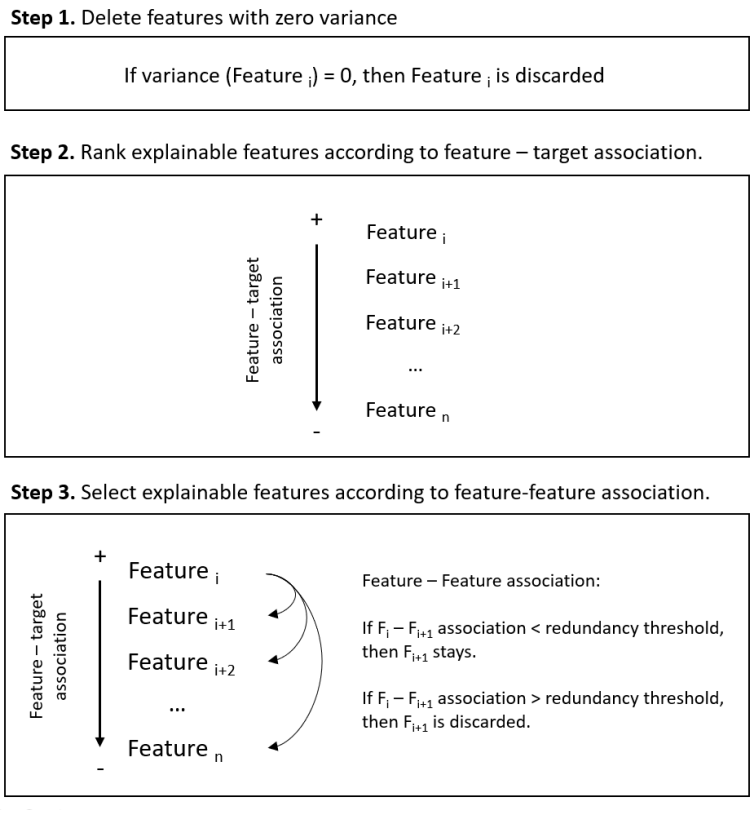


Figure 18 Feature redundancy minimisation algorithm. The minimisation of feature redundancy within a dataset can be achieved in 3 steps. **Step 1** Explanatory features with zero variance are discarded. **Step 2** Explanatory features are ranked according to their association with the target variable. Feature-target association is measured as the mean F1 score of a 3-fold cross validation routine, under the original QSPR modelling framework (estimator and hyperparameters). **Step 3** Explanatory features are selected according to their association with other features and a predefined redundancy threshold. Feature-feature association is measured using the normalised mutual information adjusted for chance. At this stage, if the association between two features exceeds the redundancy threshold, the feature lower in the rank is discarded and no longer considered. After iterating over all features, those that manage to hold their place in the rank form the solution set.

Given the discrete and low cardinality nature of most classification target variables, feature-target associations are measured as the performance with which the former predicts the latter. Performance values correspond to the mean F1 score of a 3-fold cross validation routine, under the original QSPR modelling framework - estimator and hyperparameters – one explanatory feature at a time. On the other hand, feature-feature associations are measured using normalised mutual information values adjusted for chance (ANMI). Mutual

information (MI) quantifies the information shared between two random variables. MI values are low when features are independent and maximised when one feature is a deterministic function of the other. MI normalisation projects values onto the intuitive interval [0, 1],[283] and adjustment for chance guarantees that the expected MI value obtained by chance is equal to zero.[284] MI has already been successfully used for feature selection[285–287] and redundancy minimisation.[282] The choice of this measure is often motivated by its ability to accommodate both continuous and discrete variables[288] as well as to quantify more general and complex feature relationships that other popular association metrics would miss. The nature of the computation requires continuous features to be discretised. Thus, any feature with a cardinality above 20, *i.e.* that could adopt more than 20 unique values, is discretised into 20 bins. Note that this way, feature – target and feature – feature associations are measured in the same scale and follow the same general meaning.

Although redundancy minimisation is achieved through feature association metrics, correlation does not imply redundancy.[259] Unlike most redundancy-based methods that are based on univariate strategies, our approach acknowledges the fact that the relevance and redundancy character of individual features vary according to context. Thus, critical factors such as feature interdependencies and the modelling framework must be factored in during the assessment. Accordingly, our pipeline conducts redundancy studies to understand the optimal level of redundancy that a dataset must possess for a given QSPR modelling task. During these studies, the above procedure (**Figure 18**) is repeated for different redundancy thresholds, representing the maximum association value that can exist between any two explanatory features in the solution set. We suggest the thresholds values of 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99. Each resulting feature subset is evaluated for its predictive power, measured as the mean F1 score of a 3-fold cross validation routine, once again under the original QSPR modelling framework. Results are delivered as a plot of

model performance against redundancy thresholds. The optimal level of redundancy corresponds to the lowest threshold value that enables a performance on a par with the original predictive power attained during the internal model assessment. Note that redundancy studies should be based on datasets that are cleaned, treated for outliers, scaled, and class balanced with respect to the target variable. Visit ***Chapter 2 / Methods / Data Processing*** for details on dataset processing measures.

Finally, the redundancy level is manually selected and the corresponding subset of features is returned. Note that under this newly proposed framework, the specification of the number of output features in the subset is not required, as it should not have to be. At this point, the dataset meets the assumption of feature independence, *i.e.* minimal collinearity. During this final step, the groups of redundant features identified for the predefined threshold are also compiled in a dataset redundancy report. Note that the initial ordering of features by their strength of association with the target variable was adopted as a meaningful way to reduce dataset redundancy, which in turn also defines the optimal feature space for incoming feature selection methods to operate in.

The second stage of the pipeline addresses feature relevance and it was designed to identify the subset of features that are critical to the prediction task. Generally, wrapper methods are good choices for this type of assessment. Amongst wrappers, the genetic algorithm (GA) is a stochastic tool that selects features based on the evolution of populations.[289–292] Our GA implementation is schematically represented in **Figure 19**.

In the GA context, each algorithm's iteration round is referred to as a generation, each subset of explanatory features is referred to as an individual or member, and the collection of all individuals of a given iteration round represents the population of that generation. In short, an initial collection of feature subsets is defined at random (**Figure 19,**

Step 1) and then evolved through a number of generations (**Figure 19, Steps 2 - 5**). During the first half of generations, the random search stage, a full exploration of the descriptor space is allowed. The second half is taken over by a convergence stage, where the subset of descriptors yielding the highest accuracies are refined and optimised.

After creating the first population, the algorithm operates in an iterative mode. Each generation [iteration] includes four steps: fitness, selection, mutation and crossover. During the fitness step (**Figure 19, Step 2**), the fitness of each individual is assessed by a fitness function. Individuals are then ranked and selected during the second step (**Figure 19, Step 3**). The selection includes some of the best and a few of the worst individuals of the generation. The former are kept as some of the best candidate solutions, and the latter are considered to prevent the method from getting stuck in suboptimal solutions during feature space exploration. The mutation (**Figure 19, Step 4**) and crossover (**Figure 19, Step 5**) steps modify existing population members from which new ones are produced. Whilst changes promote the efficient exploration of the solution space, the best candidate of each generation is not altered to allow convergence. The algorithm runs until the results have converged or until the maximum number of generations is achieved. Upon termination, the individual that led to the best fitness score across all generations is returned as a list of feature names, considered as the critical feature set driving the prediction of the target variable. The order of the critical features returned in the solution has no particular meaning. The technical details and parameters of each step are described in the ***Chapter 5 / Methods / Genetic Algorithm*** section.

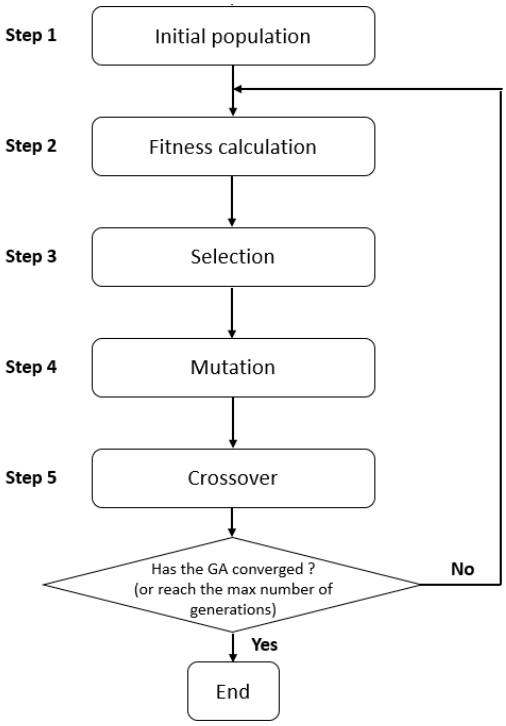


Figure 19 The genetic algorithm. The genetic algorithm starts by initialising the population, where each member represents a random sample of explanatory features. The four next steps – fitness calculation, selection, mutation, and crossover are repeated over a number of generations, until the results have converged or until the maximum number of generations is achieved. The final result consists of the subset of features that enabled the highest model performance. Refer to ***Chapter 5 / Methods / Genetic Algorithm*** for technical implementation details.

The efficient exploration of the feature space is guided by a customised fitness function that assesses each candidate solution based on its size, *i.e.* the number of features considered in the subset, and the performance that a QSPR model trained on those would enable. The performance of each subset is evaluated as the mean F1 score of a 3-fold cross validation routine in light of the modelling framework selected for the final QSPR model production. The number of critical features to be identified is unknown, but a rough guess must be passed on to the method as input. The fitness function includes a weight that controls the flexibility with which the method must stick to this exact number, by penalising the fitness of each feature subset solution based on its size. Accordingly, very low values

allow the size of the solution to expand at each generation, preventing efficient dimensionality reduction. On the other hand, very high values encourage the method to find an exact number of critical features, regardless of the true size of the optimal solution. Note that by taking the size into account, the values reported by the fitness function do not directly reflect the internal predictive power of each candidate solution.

The efficient method application requires the original array of explanatory features, the modelling framework (estimator and hyperparameters) to be used in the final QSPR model, and some training data that is cleaned, treated for outliers, scaled, and class balanced with respect to the target variable, as described in ***Chapter 2 / Methods / Data Processing***. Additionally, most GA hyperparameters are also tuneable, but default values tend to work well in most cases.

Note that conserving the original feature space, *i.e.* feature definitions, across both stages of the feature selection pipeline becomes critical to attempt future model interpretation.[257]

5.2.1.2 Feature selection pipeline implementation

In this section we use the crystallizability dataset to illustrate the step by step implementation of our feature selection pipeline. The crystallizability dataset was prepared as described in the first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model***. The feature selection pipeline was implemented as described throughout ***Chapter 5 / 5.2.1.1 Feature selection pipeline description***.

Redundancy studies results are shown in **Figure 20** and suggest that the redundancy of the crystallizability dataset (measured by the normalised mutual information adjusted for chance) can be decreased to a minimum level of 0.2 with little to no information loss to the modelling task. Below this level, decreasing redundancy starts to have some impact on the predictive power enabled by remaining features and above it, improvements in performance

do not justify the increase of the problem's dimensionality (see ***Supplementary Information K*** for more details).

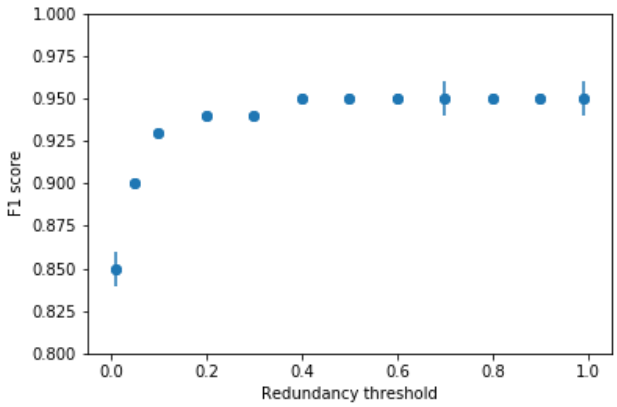


Figure 20 Redundancy study results for the crystallizability dataset. The redundancy level or threshold represents the maximum level of association allowed between any two features in the final dataset, measured as the normalised mutual information adjusted by chance. The redundancy threshold values adopted were 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99. Dimensionality corresponds to the number of features in each dataset. F1 scores represent the internal performance of models built from the crystallizability dataset containing different levels of redundancy. These where routinely calculated as the mean of the F1 score values produced from a 3-fold cross validation routine under the original crystallizability modelling framework. Dataset dimensionality details for each redundancy level can be found in ***Supplementary Information K***.

The feature selection pipeline enabled two additional versions of the crystallizability dataset: the minimally redundant version obtained from redundancy studies and the consequent maximum relevance version of that, comprising the set of critical features identified by the genetic algorithm. Each version was assessed in terms of dimensionality and predictive power, as described in ***Chapter 5 / Methods / Dimensionality Reduction Performance***. Results can be found in the **Table 13**, below.

Table 13 Evolution of the crystallizability dataset's dimensionality and predictive power, as data flows through the feature selection pipeline. The original feature set corresponds to the initial collection of features considered for QSPR modelling. The non-redundant feature set corresponds to the collection of features returned after minimising dataset redundancy (**Figure 17**, step 1). The set of critical features corresponds to the collection of features returned by the genetic algorithm (**Figure 17**, step 2).

Feature set	Dimensionality	Internal model performance
Original	113	0.95 +/- 0.02
Non-redundant	64	0.94 +/- 0.00
Critical	9	0.89 +/- 0.00

Addressing redundancy enabled a dimensionality drop from 113 to 64 features with no virtual loss of information content to the dataset. Redundancy studies further suggest that the relevance of features across the non-redundant set may vary. This is particularly noticeable in cases where no changes to the predictive power of the dataset occur, despite its increase in redundancy, *i.e.* dimensionality. In fact, redundancy has no direct impact on relevance: uncorrelated features may not be relevant, whilst highly correlated ones could all be found critical to the prediction task. In this case, performances could only be maximised when high redundancy levels are allowed in the dataset. Here, an alternative and more efficient redundancy filtering approach should consist of investigating the newly considered features, every time the increase of the redundancy threshold leads to an increase of the model's predictive power.

Addressing the relevance of non-redundant features enabled a further drop in dimensionality reduction from 64 to 9, with a cost of a 5% drop in performance. Recall that in this second step, the genetic algorithm identifies critical features as the set that optimises the trade-off between feature number minimisation and predictive capacity maximisation. Accordingly, we speculate that the overall decrease in performance from 95% to 89% justifies the extent of the dimensionality reduction achieved. Note that this trade-off is meant to

favour the speed of training and interpretability of the model, and can be tuned via the genetic algorithm parameters.

The most relevant features are listed in **Table 14**, and mostly refer to molecular size and flexibility. These findings agree with the original crystallizability prediction results reported by *Wicker et al*.[32] Although the descriptors found may have been different, both sets seem to cover the same structural aspects.

Table 14 Top most relevant features driving the crystallizability prediction problem.

Feature name	*Description*
MolWt	Molecular Weight.
FractionCSP3	Fraction of sp3 hybridised carbons.
NumRotatableBonds	Number of rotatable bonds in the molecule.
NumAliphaticHeterocycles	Number of Aliphatic Heterocycle rings.
qed	Quantitative estimation of drug-likeness.
SMR_VSA3	MOE-type descriptors using Wildman-Crippen molecular refractivity and surface area contributions. Measure of polarizability.
EState_VSA8	MOE-type descriptors using Estate indices and surface area contributions.
SlogP_VSA10	MOE-type descriptors using Wildman-Crippen log P and surface area contributions. Measure of hydrophilic/hydrophobic effects.
FpDensityMorgan2	Density of a radius 2 Morgan fingerprint

The same analysis was conducted for the HBND, crystal habit and polymorphism case studies. The corresponding datasets used in this experiment were those produced and/or described in ***Chapter 2 / 2.2.3 The models.*** Feature selection results can be found in ***Supplementary Information K - M***. In short, the features found to be most relevant to the prediction of HBND mainly cover molecular charge and interaction profiles. Critical features account for the availability of hydrogen bond interacting groups and atoms (*eg. NHOHCount, NOCount*), as well as for the distribution of charge and electronic properties across molecular surfaces (*eg. PEOE_VSA4, Estate_VSA2*), which seem to play a key role in establishing

hydrogen contacts and stabilising intermolecular interactions. On the other hand, crystal habit prediction seems to be mainly driven by features that describe the distribution of a range of molecular properties across the molecular surface. These include charge (*eg. PEOE_VSA4*), log P (*eg. SlogP_VSA3*), and molecular refractivity (*eg. SMR_VSA10*). Other structural features reported cover different aspects of the molecular shape (*eg. BalabanJ, Kappa3, Chi4n*). Finally, the prediction of whether a molecule is polymorphic seems to focus on information related to molecular size and shape (*eg. MolWt, Ipc*), as well as the charge/interaction profile of molecules. The latter is covered by features aimed to describe the topology and electronic distribution across molecular surfaces (*eg. PEOE_VSA11, Estate_VSA2*), and the potential to establish hydrogen contacts (*eg. NumHDonors, NumHAcceptors*).

The feature selection pipeline was consistently effective across all case studies. In summary, the preservation of the internal model performance after the first step confirms that only redundant information was discarded from the dataset, whilst the small drop in the second step reassures us that the most relevant features were successfully identified. In conclusion, our findings suggest the effectiveness with which our feature selection pipeline addresses the redundancy and relevance of datasets.

5.2.2 Model Interpretability

5.2.2.1 Model interpretability pipeline description

QSPR models have been mostly applied to the *in-silico* estimation of molecular properties. However, the prediction process itself may hold extremely valuable insights about structural - property relationships, which could contribute to advancing our understanding of chemical systems. Here we describe a simple but effective general-purpose model interpretability pipeline, based on standard model-agnostic methods for knowledge

extraction, *i.e.* variable importance and sensitivity studies. The system comprises three steps: data assurance, magnitude and directionality studies, and alternative feature studies, which are summarised in **Figure 21**. The major novelties of the pipeline correspond to the first and last steps, making it particular useful for the explanation of classifiers built from datasets whose features are highly correlated and whose their meanings span across different levels of abstraction.

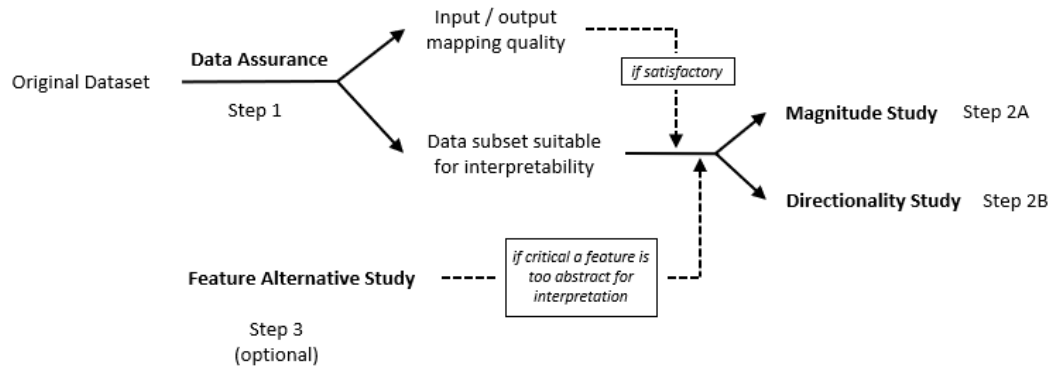


Figure 21 The model interpretability pipeline. **Step 1** During data assurance, the original dataset is narrowed down to the subset of data that reflects the prediction mechanism of the model. That includes the collection of examples found to belong to the model's applicability domain, solely represented by the critical features that drive the prediction task. The quality of the insights that can be derived from the model (input/output mapping) is also estimated. The remaining steps are only carried out if the user finds this estimate satisfactory. **Step 2** Magnitude and directionality studies are conducted for each critical feature, based on the values assumed by the instances considered at this stage. Magnitude studies report the relative importance of each feature, whilst directionality studies reveal general directionality trends. **Step 3** Feature alternative studies are conducted when critical features are too abstract for interpretation. Under these circumstances, the feature alternative study might suggest suitable and more interpretable replacements.

The data assurance step (**Figure 21, Step 1**) guarantees that the input data meets the requirements for model interpretability and estimates the quality of the insights that can be derived during these studies.

The former aim of the data assurance step refers to narrowing down the original data to the subset of critical features that drive the prediction task and collection of examples

found to belong to the model's applicability domain. The method assumes that feature selection was carried out beforehand using the strategy described in ***Chapter 5 / 5.2.1.1 Feature selection pipeline description***. Furthermore, the method also accepts that examples only need to be scanned for their prediction confidence, as they should already comply with the validity and reliability dimensions of the AD, because training and test sets are produced through the random sampling of the same initial pool of instances. The data assurance step also expects training and test data inputs to be cleaned, treated for outliers, scaled, and class balanced with respect to the target variable. Visit ***Chapter 2 / Methods / Data Processing*** for details on dataset processing measures. The original training and test sets are trimmed to only include the most relevant of the explanatory features. The training data are used to produce a new version of the QSPR model with the original modelling framework, which is used to evaluate all examples under different confidence thresholds: 0.99, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, and 0.01. Prediction confidence studies follow the same assumptions, definitions and methods described in ***Chapter 4 / 4.2.3.1 Confidence tool description***. The goal is to identify the highest threshold that enables enough confident predictions from each class, such that meaningful statistics and model insights can be derived from this data subset. The highest confidence threshold that enables at least 100 confident predictions per class is manually selected and the collection of examples predicted with confidence is returned. Note that training examples are intentionally processed by the model. This is acceptable because no conclusions about model performance are derived; we simply aim to identify all examples of the applicability domain that seem to follow the data patterns by which the model is governed.

The latter aim of the data assurance step refers to evaluating whether the prediction process can be converted into trustworthy chemical knowledge. Examples predicted with high confidence are assumed to follow most of the data patterns that drive the prediction

task, offering no doubt as to which class they belong to. Intuitively, a high number of confident and correct answers suggests that the model was able to establish the right mapping between features and target variable, whilst a high number of confident but incorrect predictions suggests otherwise. Therefore, the fraction of confident predictions that are also correct can be interpreted as a quality measure of how well the model was successfully able to map inputs to outputs and in turn, a measure of trust for any chemical insight derived from it. This measure is also reported. When the fraction of correct and confident predictions is high enough (a subjective choice to be made by the user), model interpretability studies are justified.

Magnitude studies (**Figure 21, Step 2A**) assess the magnitude of the influence that each feature has on the prediction task. In this study, the impact of each feature is measured by how well it is able to predict the target variable, represented by the mean F1 score of a 3-fold cross validation routine under the original model base estimator and parameters. Features are then sorted in descending order of magnitude and results are returned as a bar plot. The findings are used under the assumption that changes to the target variable are more likely and easily achieved by manipulating top ranked features. One of the limitations of this procedure is that no explanatory feature dependencies are considered.

Directionality studies (**Figure 21, Step 2B**) assess the direction of the influence that each feature has on the prediction task, *i.e.* the way the predicted outcome is affected when the value of each descriptor increases or decreases. These studies are conducted on an unscaled version of the subset of the original dataset identified during the data assurance step. In this step, the values of each feature are first grouped by target class labels. A probability density function is produced for each set of values via kernel density estimation, using the Gaussian kernel and a bandwidth that is automatically determined using the Scott's method.[293] probability density functions are then evaluated across the feature's domain. A

schematic representation of what the resulting class probability density profiles might look like is shown in **Figure 22**. These are useful to identify the general directionality trend of each feature and to explain specific cases that do not comply with it. In this implementation, directionality trend analysis only consider the feature values that maximise the probability density function of each class. In short, these are sorted according to the class target labels each time, for consistency. These collections of class representing values are tested for monotonicity, and their monotonic nature determines the trend or directionality of influence of the feature. In **Figure 22** for example, the class probability density function profiles suggest that larger feature values tend to be more frequent in class 1 compared to class 0. Consequently, the directionality trend derived will suggest that increasing the values of the feature might increase the likelihood of molecules to fall into class 1. Note that for multi-class problems, if labels possess no natural order, the automatic identification of trends can be compromised. The classes defined for crystal habit prediction - block, plate and needle – are an example of this. In these cases, a manual analysis should be conducted. Finally, the effect size is computed as a measure of class separation and used to evaluate the significance of each trend. That corresponds to the largest difference between the arguments of the probability density function peak values amongst all classes, normalised by the feature range. Ranges are defined between the 1^{th} and the 99^{th} percentiles to avoid any remaining outliers. These measures are schematically represented in **Figure 22**. Intuitively, for a constant feature range or domain, closer class probability density function peaks suggest a lower ability of the feature to discriminate between classes, which is reflected in smaller effect size values. At this stage, trends are considered significant if the effect size, i.e. the largest peak difference was at least 5% of the corresponding feature range, meaning that the peaks between (at least) two classes are distinct. The effect size threshold is a subjective choice to be made by the user. On the other hand, any feature that shows no significant effect size is no longer

considered critical to the prediction task, as there is no clear evidence that it alone contributes to the separation of classes. This judgement must be a) supported by the findings of magnitude studies, where such features can be found to individually predict the target variable with a performance close to the random benchmark, and b) proven by the re-evaluation of the predictive power of the remaining features.

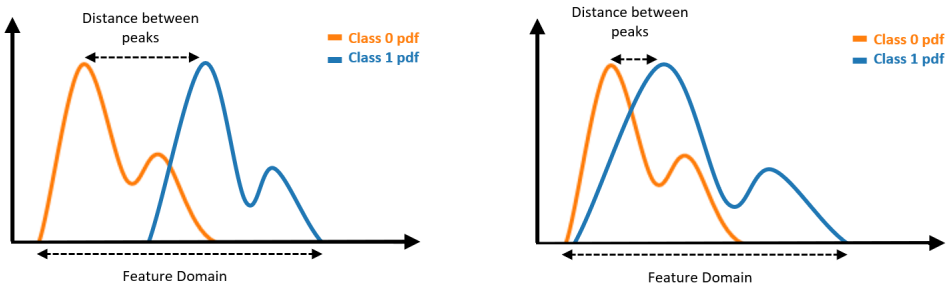


Figure 22 Schematic representation of class probability density function profiles of a hypothetical feature. Class probability density functions are computed from the group of feature values separated by target class, using the Gaussian kernel and a bandwidth determined using the Scott's method. Class probability density functions are evaluated across the entire feature domain. The distance between peaks normalised by the feature domain (defined between the 1^{th} and the 99^{th} percentiles) corresponds to the effect size. This measures class separation and it is used to evaluate the significance of the directionality trend of the corresponding feature. The two examples above convey the idea that, for a constant feature range or domain, closer class probability density function peaks suggest a lower ability of the feature to discriminate between classes.

Alternative feature studies (**Figure 21, Step 3**) provide descriptor alternatives to those that were considered critical to the prediction task but are too abstract to enable a clear interpretation of how molecular structures influence the prediction of a target property. Alternative features are identified based on feature redundancy and assumed to induce similar but weaker influences on the prediction task, when compared to the original option. The method takes in the set of most relevant features and the dataset redundancy report produced during feature selection to generate a series of bar plots. Each plot

corresponds to a critical feature and displays its suitable replacements sorted by their level of association with the target variable. From the strongest to the weakest level of association, these are considered the next best option for feature substitution. The selection of replacements is left to the user and must be focused on maximising the interpretability of the system. If a feature is to be replaced, its magnitude and directionality are also calculated as described above. Note that the feasibility of this step is determined by the implementation of our feature selection pipeline, and its success is dependent on the redundancy contained in the original dataset.

5.2.2.2 Model interpretability pipeline implementation

In this section we use the crystallizability case study to illustrate the step by step implementation of our model interpretability pipeline. Initial inputs to the system include the original dataset, the corresponding redundancy report, and the collection of most relevant features found to drive the prediction task. The crystallizability original dataset was prepared as described in the first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model***. The last two items were retrieved from the earlier feature selection pipeline implementation. The model interpretability pipeline was adopted as described throughout ***Chapter 5 / 5.2.2.1 Model interpretability pipeline description***.

First, the crystallizability training and test sets were randomly sampled from the original dataset to accelerate the runtime of the procedure. A new crystallizability model was prepared under the original modelling framework, and used throughout this section. **Table 15** displays the total number of confident examples and the fraction of those predicted correctly, when input data prediction is constrained by different confidence thresholds.

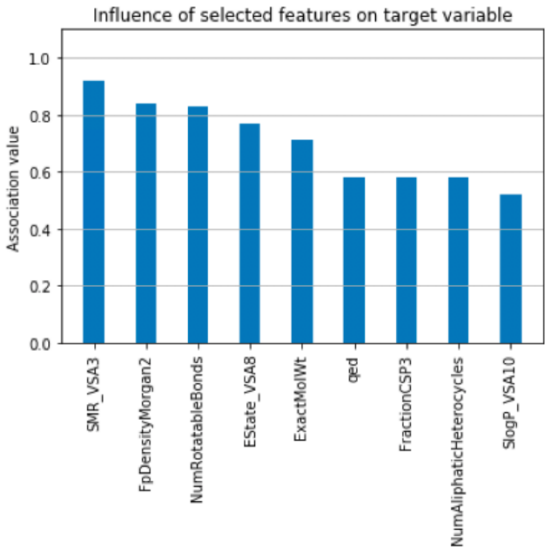
Table 15 Data assurance results for the crystallizability case study. Refer to ***Chapter 5 /5. 2.1.1 Feature selection pipeline description*** for details.

Confidence Threshold	Confident predictions	Fraction of correct predictions
0.99	20	0.99
0.90	1584	0.98
0.80	4172	0.97
0.70	5740	0.96
0.60	6968	0.95
0.50	7676	0.93
0.40	8186	0.92
0.30	8634	0.91
0.20	9106	0.90
0.10	9478	0.90
0.01	9814	0.89

Based on these findings, the data were narrowed down to the group of examples that could be predicted with a confidence threshold of at least 0.90. The new subset includes 1584 confident examples equally distributed between both classes, out of which 98% also correspond to correct predictions. This sets the basis for model interpretability, where the high volume of confident examples and large fraction of correct predictions should suffice to robustly derive meaningful and trustworthy conclusions about the crystallizability prediction procedure during the following steps. The same analysis was carried out for the HBND, crystal habit and polymorphism datasets. As expected, the fraction of confident and correct predictions obtained for any of our limited performing QSPR classifiers was considered too low to attempt any knowledge extraction. Out of all HBND, crystal habit and polymorphism examples that could be predicted with the highest confidence, only 51%, 70% and 71% led to correct answers. These findings are consistent with the low quality of their feature-property mapping task. Under these circumstances, model interpretability studies were discontinued as the results could be misinterpreted.

Magnitude study results are shown in **Figure 23**. The results suggest that desired changes in the outcome are most likely and easily achieved through the manipulation of structural features such as molecular polarizability (*SMR_VSA3*) and or number of rotatable bonds (*NumRotatableBonds*). On the other hand, features like drug-likeness (*qed*), the number of aliphatic heterocycles in the molecule (*NumAliphaticHeterocycles*) and its hydrophilic/hydrophobic character (*Slogp_VSA10*) seem to have the lowest control over the target variable, and because of their complicated relationship with molecular structures, changes through their manipulation may be more challenging to achieve.

Figure 23 Magnitude study results for the crystallizability case study. Feature-target association is measured as the mean F1 score of a 3-fold cross validation routine, under the original QSPR modelling framework (estimator and hyperparameters). Refer to ***Chapter 5 / 5.2.1.1 Feature selection pipeline description*** for details.



Directionality study results are shown in **Table 16**. First, note that feature *SlogP_VSA10* was discarded for not showing a significant effect size. Indeed, the feature was

shown to exhibit a lower magnitude of impact and its withdrawal seems to produce little to no loss of predictive power of the remaining set. We speculate that such cases occur when the chosen number of critical features to be found by the genetic algorithm step is larger than the true number of critical features, or because those structural aspects are indirectly covered by other features. The directionality trends towards the crystallizable class are revealed for the remaining features. In short, the findings suggest that generally decreasing the values of any of those features makes molecules more prone to be assigned to the crystallizable class.

Table 16 Directionality study results for the crystallizability case study. Directionality trends suggest how feature values must be altered, such that the likelihood of a molecule falling into a given target class is increased. The trends presented in the table below are those that increase the likelihood of molecules to fall into the crystallizable class. Refer to ***Chapter 5 / 5.2.2.1 Model interpretability pipeline description*** for details on directionality studies and trend identification.

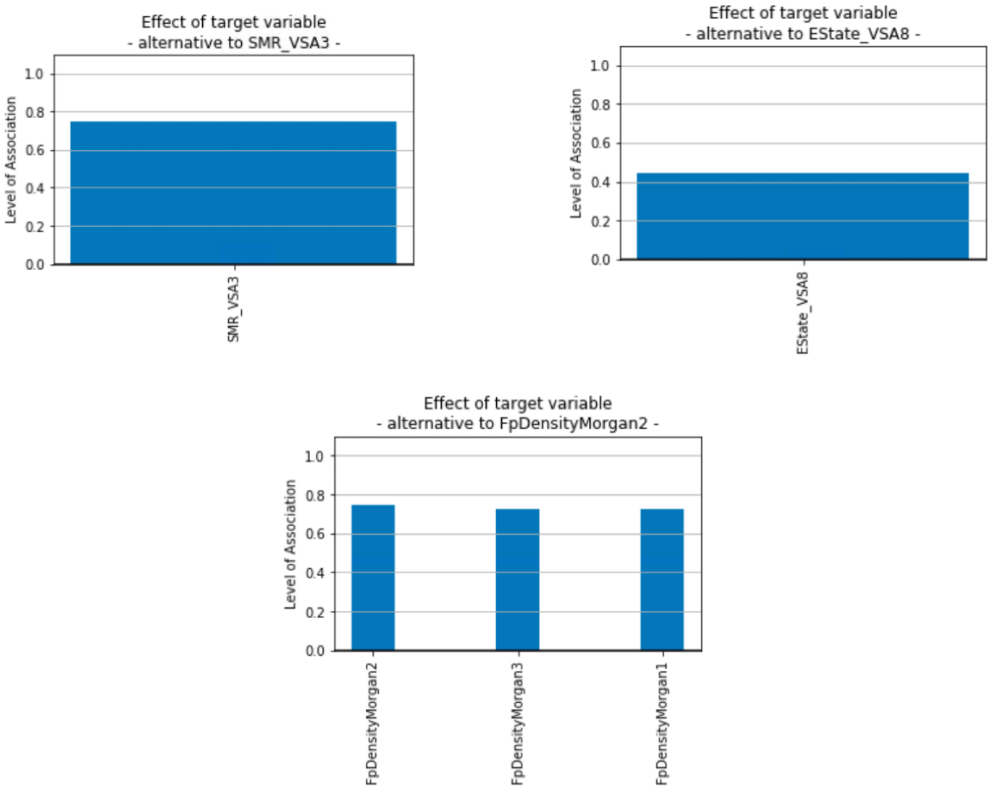
Critical Feature	Directionality Trend
SMR_VSA3	Decrease
FpDensityMorgan2	Decrease
NumRotatableBonds	Decrease
qed	Decrease
FractionCSP3	Decrease
NumAliphaticHeterocycles	Decrease
MolWt	Decrease
Estate_VSA8	Decrease

These continue to mostly cover molecular size and flexibility. Furthermore, these trends comply with the thermodynamic arguments adopted by *Wicker et al*.[32] to rationalise the importance and influence of such structural properties on the prediction of crystallization propensity. In short, the decrease of flexibility factors such as the fraction of sp^3 hybridised carbons, the number of rotatable bonds, or the number of aliphatic heterocycles leads to the

existence of fewer conformers in solution, allowing those that are in the correct conformation to crystallize to occur in higher concentrations.[32] Increasing the concentration of molecules in the correct conformers near the crystal/solution boundary favours the crystallization process.[294] Similarly, the decrease of molecular size or volume factors such as those captured by descriptors for molecular weight, drug-likeness or the number of aliphatic heterocycles has been used to explain increased crystallizability due to the relative ease of rearranging solvent around smaller crystallizing molecules during attachment to a crystal growth surface.[32] Finally, features such as *Estate_VSA8* and *SMR_VSA3* are known to cover aspects related to the distribution of molecular charge across molecular surfaces, but their unintuitive meaning prevents us from a simple feature-property relationship rationalisation. Lastly, and given the fact that we do not follow well established causal inference procedures, we emphasize that no causal relationships between structural feature and property values are claimed.

The level of abstraction of some descriptors represents a challenge to the effective manipulation of molecular structures towards the production of the desired effect in the outcome. Feature alternative studies aim to solve this problem. Alternative features to *Estate_VSA8, SMR_VSA3,* and *FpDensityMorgan2* are displayed in **Figure 24**. Bar plots are expected to include a selection of alternatives ranked by their magnitude of association with the target variable. From left to right, they are assumed to represent the next best replacement. Unfortunately, no feature alternatives were available for *Estate_VSA8 or SMR_VSA3.* Furthermore, those available for *FpDensityMorgan2* were equally abstract. These studies are limited to the redundancy contained in the original dataset, and no benefits were added by this step to the maximisation of the interpretability of the crystallizability predictive model.

Figure 24 Feature alternative study results for the crystallizability prediction problem. Feature-target association is measured as the mean F1 score of a 3-fold cross validation routine, under the original QSPR modelling framework (estimator and hyperparameters). Refer to ***Chapter 5 / 5.2.1.1 Feature selection pipeline description*** for details.



Despite the unfortunate fact that feature alternative studies could not add value to the interpretation of the crystallizability prediction task, it is still worth illustrating how this optional step could help maximise knowledge extraction from QSPR models. Hence, we showcase some of the results that the feature alternative study would have enabled in the HBND case (**Figure 25**). Suppose the count of NH and OH groups constitutes critical information for the prediction of HBND. Furthermore, a practitioner would like to manipulate the molecular structure of a given compound based on this descriptor, to increase the likelihood of the new molecule to fall in the desired target class. This second-best alternative

could help the practitioner rationalise that the importance of the number of NH and OH groups for the prediction of HBND might be related to the availability of hydrogen donor groups in the molecule. If the number of NH or OH groups cannot be altered, alternative changes can be considered with affect only the alternative features. In this case knowing that the number of hydrogen bond donors is also correlated, we could add or remove less common donor groups such as CF_2H,[295] S-H or P-H groups.[296] Feature alternative results can help accelerate this type of thought process and strengthen the foundations of new structure-property relationship hypothesis and interpretations. Note that based on data assurance results for the HBND case, no knowledge, insights, or conclusions are derived from this illustrative example.

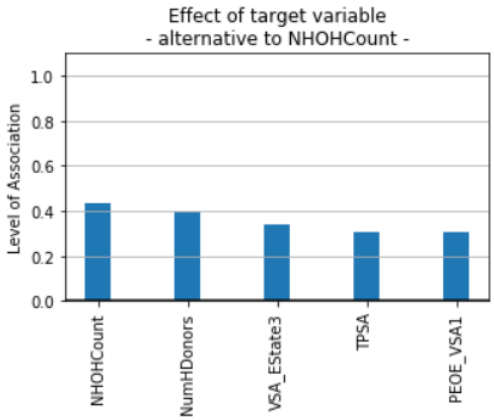
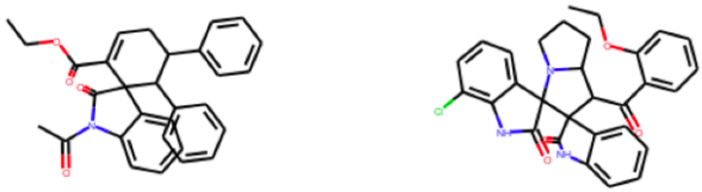


Figure 25 Feature alternative study results for the HBND prediction problem – an illustrative example of how feature alternative studies could add value to the interpretation of QSPR models.

Directionality trends were validated by comparing pairs of structurally similar molecules that belong to opposite classes, as described in ***Chapter 5 / Methods / Knowledge Extraction Validation***. Structural similarities were computed based on the molecular

fingerprints of each molecule and the Tanimoto score. Here we define the change effect, which assesses whether the differences between the feature values of the examples are significant enough to consider that the trend has been satisfied or contradicted, based on a predefined threshold. We define change effect as the absolute difference between the feature values of each molecule, normalised by the feature range. Only the range between the 1th and the 99th percentiles was considered to avoid any remaining outliers. **Figure 26** displays an example of a pair of molecules used to validate the trends. Although molecules share a high Tanimoto score of 0.81, one crystallizes and the other has not yet been found to crystallize. Due to their similarities, some feature values are expected to remain constant or be very alike. Thus, we only focus on features for which the change effect was higher than 0.05. Results are summarised in **Table 17**. Essentially, this assumes that the number of rotatable bonds, drug-likeness and *Estate_VSA8* character is essentially the same in both molecules. The remaining five structural feature values seem to comply with the expected trends from the non-crystallizable to the crystallizable molecule.



Molecule A **Molecule B**

Figure 26 Molecular diagrams of the Molecule A (non crystallizable class) and Molecule B (crystallizable class) pair example. **Molecule A** SMILES notation: CCOc1ccccc1C(=O)C1C2CCCN2C2(C(=O)Nc3c(Cl)cccc23)C21C(=O)Nc1ccccc21. **Molecule B** SMILES notation: CCOC(=O)C1=CCC(C(c2ccccc2)C21C(=O)N(C(C)=O)c1ccccc21)c1ccccc1. Diagrams were obtained from the CSD.

Table 17 Validation of the direction of the influence of each descriptor applied to the Molecule A and Molecule B pair example. Expected trends are derived from the Directionality Study (see ***Chapter 5 / 5.2.2.1 Model interpretability pipeline description*** for details). The change effect is measured as the absolute difference between the feature values of each molecule, normalised by the feature range. A predefined threshold of 0.05 was used to assess whether the difference of feature values between both molecules were significant. Refer to ***Chapter 5 / Methods / Knowledge Extraction Validation*** for implementation details.

Descriptor	Expected behaviour	Mol A	Mo B	Change Effect
SMR_VSA3	Decrease	4.90	0	0.16
FpDensityMorgan2	Decrease	1.76	1.51	0.12
NumRotatableBonds	Decrease	4	4	**0**
qed	Decrease	0.48	0.49	**0.02**
FractionCSP3	Decrease	0.3	0.23	0.07
NumAliphaticHeterocycles	Decrease	4	1	0.75
MolWt	Decrease	527.16	465.19	0.07
Estate_VSA8	Decrease	15.53	12.13	**0.02**

Following the implementation of the model interpretability pipeline and validation of directionality trends, we comment on the limitations of each step. Magnitude studies are univariate analyses that do not consider feature interdependencies, and directionality studies derive trends from single class probability density function summary measures, both leading to the loss of important information that could enable a more accurate knowledge extraction. Finally, one of our aims was to illustrate the value of two new proposed steps for model-agnostic interpretation pipelines: data assurance and feature alternative studies. The data assurance step efficiently guarantees that the input data meets the requirements for model interpretability. Yet, the quality of the knowledge that can be extracted from QSPR models is effectively determined by the accuracy of the mapping task. In fact, models with limited predictive power were already shown to produce questionable insights.[63,69] Lastly, the feature alternative studies were not able to provide better descriptor options for the crystallizability case. Again, note that the success of this step is dependent on the redundancy

contained in the original dataset. Given these limitations, the knowledge derived from the crystallizability case study requires further confirmation and must be used with care.

5.3 Conclusion

Molecular descriptors are designed to highlight particular features of chemical structures. During the study of unclear structure-property relationships, it becomes tempting to consider as many features as possible to achieve better levels of structural description. However, blindly expanding the dimensionality of data carries several drawbacks that can easily lead to model deterioration. Furthermore, molecular descriptors are often derived from one another, which promotes the multicollinearity between features and the redundancy of the dataset content. Finally, the meaning of these descriptors span across different levels of interpretability, where a large fraction is so abstract that no valuable chemical insights can be learnt from them. QSPR approaches try to overcome these common chemical dataset characteristics by combining traditional machine learning techniques with feature selection and model interpretability strategies which help to avoid high dimensional data problems, whilst increasing the performance and interpretability of models.[251,297]

In this section we provide our contribution to the field by reporting the feature selection and the model interpretability pipelines. The feature selection pipeline operates under a minimum-redundancy maximum-relevance framework to identify the set of most relevant and minimally correlated features for a given QSPR modelling task. Our pipeline takes a new and powerful approach to the evaluation of feature redundancy and relevance, shown to effectively alleviate the curse of dimensionality and its consequent drawbacks with no information loss to the system or model performance deterioration. The model interpretability pipeline is a simple and model-agnostic solution designed to investigate how each feature influences the response variable in terms of magnitude and directionality. This

strategy extracts chemical insights from data that can be used to guide the virtual screening and optimisation of molecules, or even set the basis for new mechanistic hypothesis and theories. The design, assumptions and limitations of each method were discussed. Step by step pipeline implementations were also provided.

Our studies demonstrate that given the right tools, it becomes possible to overcome the curse of dimensionality and a new way to advance our chemical understanding is also unlocked. Unfortunately, the latter continues to be restricted to datasets with high predictive power.

5.4 Methods

Genetic Algorithm

The genetic algorithm (GA) operates on a given training set that is assumed to be cleaned, treated for outliers, scaled, and class balanced with respect to the target variable. Visit ***Chapter 2 / Methods / Data Processing*** for details on dataset processing measures. Other inputs include the original list of feature names and the original modelling framework – estimator and hyperparameters. The remaining GA parameters can also be tuned, but default values should enable good results: early_stopping=True, whether to stop the algorithm if it converges before all iterations have taken place; d=1, parameter that controls the penalty of the number of features in the fitness function; n_generations=1000, number of total iterations; n_random_search_generations= 500, number of iterations dedicated to the random search stage; n_individuals=10, number of members per generation; n_parents=5, number of members to consider for crossover; n_selected_features=10, rough number of features to be selected; mutation_threshold_RandomSearch=0.3, mutation threshold adopted during the random stage; mutation_threshold_Convergency=0.9,

mutation threshold adopted during the convergence stage; best_parents_percent=0.8, fraction of best performing parents to consider for mating.

In this context, individuals represent different subsets of features from the original dataset. The first population was initialised with 10 individuals of size 10, selected at random. The population size was fixed at 10 members per generation and the algorithm was run for 1000 generations. Half of the generations were allocated to the random search stage and the other half comprised the convergence stage.

Each generation included four steps: fitness, selection, mutation and crossover. In the fitness step, the performance of each individual is assessed by the fitness function, which considers the number of features and the predictive power of the corresponding subset of the original data matrix. Note that the fitness function contains a weight that controls the solution size penalty, which becomes larger with the increase in the number of features. The predictive power is computed as a mean F1 score of a 3-fold cross validation routine under the original modelling framework. In the selection step, individuals are ranked according to their fitness and selected for the next step: the best 4 and the single worst member. Considering low fitness individuals enables a better exploration of the parameter space and avoids the method to get stuck in local minima. In the mutation step, individuals are converted into binary vectors of length equal to the number of total descriptors in the dataset, such that each index unequivocally corresponds to one feature. In each vector, ones were assigned to indices for which the correspondent features were included in the subset; zeros were assigned otherwise. During this step, selected individuals are changed with a mutation rate of 70% (random search stage) or 10% (convergence stage) per vector element, allowing a diverse set of feature combinations to be formed and tested. The best individual of each generation was never subject to mutation, allowing the algorithm to converge. Mutation was implemented such that the number of features represented by each individual

could vary slightly but would remain relatively low, around 5-15 features. All binary vectors are then converted back into lists of feature indices. During the crossover step, 5 new individuals were produced at each generation. Here, all existing members have equal chance of participating in the mating operation, and the new child would randomly receive half of the chromosome of each parent. Upon termination, the subset of features that enabled the highest fitness across all generations is returned.

Dimensionality Reduction Performance

Dimensionality reduction performance was assessed in terms of dataset's dimensionality and predictive power. The dimensionality corresponds to the number of independent variables in the dataset. The predictive power was measured as the mean F1 score of a 3-fold cross validation routine under the original modelling framework – estimator and hyperparameters.

Knowledge Extraction Validation

The direction of influence of each descriptor was assumed to correspond to the monotonic trend that the class representing values follow (increase or decrease) from the non-crystallizable to the crystallizable class. Trend validation was done by identifying pairs of structurally similar molecules belonging to different classes, and checking whether these would follow the trend set for each descriptor. Molecular fingerprints were calculated for all crystallizable and non-crystallizable labelled molecules within the model's applicability domain, using RDKit. Structural similarity was based on fingerprint similarity, also calculated using RDKit. Only pairs of molecules whose Tanimoto similarity score was above 0.8 were considered. The feature values of each molecule were retrieved for trend validation. The change effect was calculated as the absolute difference between the feature values of each

molecule, normalised by the feature range. Only the range between the 1^{th} and the 99^{th} percentiles was considered to avoid any remaining outliers. Changes in feature values were considered significant when the change effect was equal or larger than 5%. Only features complying with a significant change effect were analysed, and trend validation was based on how many of those followed their expected trends.

Chapter 6 - Conclusion

Modern chemoinformatics relies heavily on data driven approaches. This has led to great progress in the last few decades powered by the fast-growing amount of publicly available data, advances in modelling procedures and inference techniques, and an increase in the computational power of machines. This transition has enabled computational tools in the domain of machine and deep learning that are now more powerful than ever. Indeed, many of the tools, techniques and knowledge that set the basis of chemoinformatics approaches are borrowed from active areas of research. The discipline often benefits from the advancements of its anchor fields, but it also inherits the challenges of adopting state of the art solutions. Therefore, never has chemoinformatics delivered so much impact, but also faced so many challenges to overcome and procedures to rationalise.

The disappointing results in certain applications discourage practitioners from fully embracing or expending resources on QSAR/QSPR studies, who argue that these have yet to demonstrate their robust and consistent ability to add value. The scepticism around the quality of these approaches stems for a number of reasons. Individual models often exhibit limited performance during production, lack of generalisability to unseen examples after deployment, and/or vague interpretability for knowledge extraction. In a wider context, the field faces a general lack of consistency of results and procedures. However, the potential, size, and impact that QSAR/QSPR models have demonstrated thus far has motivated the need to add structure to the discipline through the clarification of concepts, establishment of standards and policies, and provision of general-purpose open-source implemented solutions.

In this DPhil research project, I have investigated some of the Chemoinformatics challenges mentioned above, through the study of the predictive models of three emergent

properties of molecular materials. The prediction problems were selected to represent a range of model performances from ‘quite good’ to ‘only just better than random’, to demonstrate that the quality and amount of results derived from QSPR studies is often a trade-off that must be understood and considered at all times.

In the first part of this thesis we conclude that the incomplete representation of molecular structures may be one of the major factors limiting the predictive power of QSPR models to higher performances. Furthermore, improving structural descriptions through molecular feature design seems to be a powerful solution to the problem. However, it may require a long and continuous improvement for it to generate significant results. The newly developed library of intermolecular interaction propensity descriptors cleanly captured some information about crystallizability, but did not lead to measurable improvements of the performance of the predictive models of this study.

In the second half of the project, I review fundamental concepts in chemistry, statistics and machine learning to motivate, design and implement new or improved QSPR solutions. The generalisability and confidence tools address the applicability domain of models. The proposed strategies demonstrate that controlling the quality of inputs and outputs can help to identify the regions of chemical space where QSPR predictive models hold their predictive power and produce confident predictions. Their illustrative implementation alone led to performance increases of 10%, 11% and 7% of the HBND, crystal habit and polymorphism predictive models, respectively.

The feature selection and model interpretation pipelines address the curse of dimensionality and the loss of model interpretability to highly predictive features with abstract meanings. The tools were shown to efficiently reduce the dimensionality of QSPR datasets with little to no information loss to the system and the potential of maximising the interpretability and quality of the insights derived from QSPR models. For example, the

higher crystallization propensity of molecules was found to be correlated to the lower volume and flexibility of molecules.

Furthermore, all these tools constitute general-purpose solutions to the narrow scope, context dependency, or limited functionality of existing QSPR strategies. We believe that making such tools readily available to the scientific community may help to standardise the approaches of the field, add consistency to its studies, enable the direct comparison of results, and comply with policy makers, thereby adding trustworthiness, and promoting the translation of science into practical solutions for the real world.

Chapter 7 - Future Work

Following the motivations and basis for the methodologies reported in this thesis, some of the next steps that would be interesting to take concern their further development and validation. These would include:

- the validation and extension of the hydrogen bonding constants α and β of the Intermolecular Interaction Propensity library of descriptors, so more functional groups could be accounted for (***Chapter 3 / 3.2.1 Feature library implementation***). We also note that there may be better ways of parameterising electronic properties of molecules computed from density functional theory methods, which could also be pursued.
- the implementation of additional multivariate outlier detection methods for the generalisability tool (***Chapter 4 / 4.2.2.1 Generalisability tool description***)
- the automated computation of calibrated probabilities of class membership for model estimators that do not naturally produce probabilistic outcomes, for the confidence tool (***Chapter 4 / 4.2.3.1 Confidence tool description***)
- the development of the genetic algorithm, such that the size of its solution, *i.e.* the number of relevant features to be retrieved, does not have to be set by the user beforehand (***Chapter 5 / 5.2.1.1 Feature selection pipeline description***). We anticipate that this type of improvement might require analysis of a large number of datasets to ensure its effectiveness for a large range of problems.
- the assessment of the value of the data assurance and feature alternative steps of model interpretation within other existing model-agnostic pipelines (***Chapter 5 / 5.2.2.1 Model interpretability pipeline description***)

- the generalisability tool and the feature selection pipeline are two of the main methodologies of this project planned to be published in the near future.
- all the methodologies reported would also be validated with other well-studied QSPR datasets and against relevant literature.
- finally, it would be interesting to revisit the HBND, crystal habit, and polymorphism case studies adopting an extended molecular representation of compounds. This would include functional group counts, and other two and lower dimensional molecular descriptors beyond those provided by the RDKit library.
- in a wider scope, it would be of value to develop equivalent methodologies, tools and pipelines for regression problems, where target variables are continuous.

We believe that steps like the above can contribute towards the standardisation and adoption of general and well established QSPR/QSAR approaches, and the raising of awareness of the limitations of QSPR/QSAR studies.

Chapter 8 - Publications

Frade A, McCabe P, Cooper R. Increasing the performance, trustworthiness and practical value of machine learning models: a case study predicting hydrogen bond network dimensionalities from molecular diagrams. CrystEngComm. 2020;22(43):7186-7192. doi:10.1039/d0ce00111b

References

1. Varnek A, Baskin I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol Inform*. **2011**, *30*(1), 20-32.
2. Engel T. Basic Overview of Chemoinformatics. *J Chem Inf Model.* **2006**, *46*(6), 2267-2277.
3. Hansch C, Muir R, Fujita T, Maloney P. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature.* **1962**, *194*(4824), 178-180.
4. Mitchell J. Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science.* **2014**, *4*(5), 468-481.
5. Lusci A, Pollastri G, Baldi P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J Chem Inf Model.* **2013**, *53*(7), 1563-1575.
6. Lo Y, Rensi S, Torng W, Altman R. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today.* **2018**, *23*(8), 1538-1546.
7. Cherkasov A, Muratov E, Fourches D, Varnek A, Baskin I, Cronin M, Dearden J, Gramatica P, Martin Y, Todeschini R, Consonni V, Kuz'min V, Cramer R, Benigni R, Yang C, Rathman J, Terfloth L, Gasteiger J, Richard A, Tropsha A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J Med Chem.* **2014**, *57*(12), 4977-5010.
8. Polański J, Gieleciak R, Bąk A. The Comparative Molecular Surface Analysis (COMSA) − A Nongrid 3D QSAR Method by a Coupled Neural Network and PLS System: Predicting pKa Values of Benzoic and Alkanoic Acids. *J Chem Inf Comput Sci.* **2002**, *42*(2), 184-191.
9. Xing L, Glen R, Clark R. Predicting pKa by Molecular Tree Structured Fingerprints and PLS. *J Chem Inf Comput Sci.* **2003**, *43*(3), 870-879.
10. Martin D, Sild S, Maran U, Karelson M. QSPR Modeling of the Polarizability of Polyaromatic Hydrocarbons and Fullerenes. *The Journal of Physical Chemistry C.* **2008**, *112*(13), 4785-4790.
11. Basak S, Gute B, Grunwald G. Use of Topostructural, Topochemical, and Geometric Parameters in the Prediction of Vapor Pressure: A Hierarchical QSAR Approach. *J Chem Inf Comput Sci.* **1997**, *37*(4), 651-655.
12. Nigsch F, Bender A, van Buuren B, Tissen J, Nigsch E, Mitchell J. Melting Point Prediction Employing k-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J Chem Inf Model.* **2006**, *46*(6), 2412-2422.
13. Hughes L, Palmer D, Nigsch F, Mitchell J. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J Chem Inf Model.* **2008**, *48*(1), 220-232.
14. Katritzky A, Kuanar M, Slavov S, Hall C, Karelson M, Kahn I, Dobchev D. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem Rev.* **2010**, *110*(10), 5714-5789.
15. Stein S, Brown R. Estimation of normal boiling points from group contributions. *J Chem Inf Comput Sci.* **1994**, *34*(3), 581-587.
16. Katritzky A, Lobanov V, Karelson M. Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure−Property Relationship. *J Chem Inf Comput Sci.* **1998**, *38*(1), 28-41.
17. Karelson M, Perkson A. QSPR prediction of densities of organic liquids. *Comput Chem.* **1999**, *23*(1), 49-59.
18. Toropov A, Toropova A. QSPR modeling of alkanes properties based on graph of atomic orbitals. *Journal of Molecular Structure: THEOCHEM.* **2003**, *637*(1-3), 1-10.

19. Monnery W, Svrcek W, Mehrotra A. Viscosity: A critical review of practical predictive and correlative methods. *The Canadian Journal of Chemical Engineering.* **1995**, *73*(1), 3-40.
20. Ivanciuc O, Ivanciuc T, Filip P, Cabrol-Bass D. Estimation of the Liquid Viscosity of Organic Compounds with a Quantitative Structure−Property Model. *J Chem Inf Comput Sci.* **1999**, *39*(3), 515-524.
21. Erić S, Kalinić M, Popović A, Zloh M, Kuzmanovski I. Prediction of aqueous solubility of drug-like molecules using a novel algorithm for automatic adjustment of relative importance of descriptors implemented in counter-propagation artificial neural networks. *Int J Pharm.* **2012**, *437*(1-2), 232-241.
22. Louis B, Agrawal V, Khadikar P. Prediction of intrinsic solubility of generic drugs using MLR, ANN and SVM analyses. *Eur J Med Chem.* **2010**, *45*(9), 4018-4025.
23. Cao D, Xu Q, Liang Y, Chen X, Li H. Prediction of aqueous solubility of druglike organic compounds using partial least squares, back-propagation network and support vector machine. *J Chemom.* **2010**, *24*(9), 584-595.
24. Wang R, Fu Y, Lai L. A New Atom-Additive Method for Calculating Partition Coefficients. *J Chem Inf Comput Sci.* **1997**, *37*(3), 615-621.
25. Wang R, Gao Y, Lai L. Calculating partition coefficient by atom-additive method. *Perspectives in Drug Discovery and Design.* **2000**, *19*(1), 47-66.
26. Katritzky A, Sild S, Karelson M. General Quantitative Structure−Property Relationship Treatment of the Refractive Index of Organic Compounds. *J Chem Inf Comput Sci.* **1998**, *38*(5), 840-844.
27. Katritzky A, Sild S, Karelson M. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J Chem Inf Comput Sci.* **1998**, *38*(6), 1171-1176.
28. Schweitzer R, Morris J. Improved Quantitative Structure Property Relationships for the Prediction of Dielectric Constants for a Set of Diverse Compounds by Subsetting of the Data Set. *J Chem Inf Comput Sci.* **2000**, *40*(5), 1253-1261.
29. Bosque R, Sales J. Polarizabilities of Solvents from the Chemical Composition. *J Chem Inf Comput Sci.* **2002**, *42*(5), 1154-1163.
30. Huibers P, Lobanov V, Katritzky A, Shah D, Karelson M. Prediction of Critical Micelle Concentration Using a Quantitative Structure−Property Relationship Approach. 1. Nonionic Surfactants. *Langmuir.* **1996**, *12*(6), 1462-1470.
31. Roberts D. Application of Octanol/Water Partition Coefficients in Surfactant Science: A Quantitative Structure−Property Relationship for Micellization of Anionic Surfactants. *Langmuir.* **2001**, *18*(2), 345-352.
32. Wicker J, Cooper R. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm.* **2015**, *17*(9), 1927-1934.
33. Ghosh A, Louis L, Arora K, Hancock B, Krzyzaniak J, Meenan P, Nakhmanson S, Wood G. Assessment of machine learning approaches for predicting the crystallization propensity of active pharmaceutical ingredients. *CrystEngComm.* **2019**, *21*(8), 1215-1223.
34. Rupp M, Tkatchenko A, Müller K, von Lilienfeld O. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys Rev Lett.* **2012**, *108*(5), 058301.
35. Montavon G, Rupp M, Gobre V, Vazquez-Mayagoitia A, Hansen K, Tkatchenko A, Müller K, Anatole von Lilienfeld O. Machine learning of molecular electronic properties in chemical compound space. *New J Phys*. **2013**, *15*(9), 095003.
36. Hansen K, Biegler F, Ramakrishnan R, Pronobis W, von Lilienfeld O, Müller K, Tkatchenko A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J Phys Chem Lett.* **2015**, *6*(12), 2326-2331.

37. Begam B, Kumar J. A Study on Cheminformatics and its Applications on Modern Drug Discovery. *Procedia Eng.* **2012**, *38*, 1264-1275.
38. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today.* **2017**, *22*(11), 1680-1685.
39. Kubinyi, H. Success stories of computer-aided design. *Computer Applications in Pharmaceutical Research and Development*. **2006**, *2*, 377
40. Khan M, Sylte I, Khan M, Sylte I. Predictive QSAR Modeling for the Successful Predictions of the ADMET Properties of Candidate Drug Molecules. *Curr Drug Discov Technol.* **2007**, *4*(3), 141-149.
41. Worachartcheewan A, Nantasenamat C, Owasirikul W, Monnor T, Naruepantawart O, Janyapaisarn S, Prachayasittikul S, Prachayasittikul V. Insights into antioxidant activity of 1-adamantylthiopyridine analogs using multiple linear regression. *Eur J Med Chem*. **2014**, *73*, 258-264.
42. Prachayasittikul V, Worachartcheewan A, Shoombuatong W, Songtawee N, Simeon S, Nantasenamat C. Computer-Aided Drug Design of Bioactive Natural Products. *Curr Top Med Chem*. **2015**, *15*(18), 1780-1800.
43. Prachayasittikul V, Pingaew R, Worachartcheewan A, Nantasenamat C, Prachayasittikul S, Ruchirawat S. Synthesis, anticancer activity and QSAR study of 1,4-naphthoquinone derivatives. *Eur J Med Chem*. **2014**, *84*, 247-263.
44. Nantasenamat C, Li H, Mandi P, Worachartcheewan A, Monnor T, Isarankura-Na-Ayudhya C, Prachayasittikul V. Exploring the chemical space of aromatase inhibitors. *Mol Divers*. **2013**, *17*(4), 661-677.
45. Nabu S, Nantasenamat C, Owasirikul W, Lawung R, Isarankura-Na-Ayudhya C, Lapins M, Wikberg J, Prachayasittikul V. Proteochemometric model for predicting the inhibition of penicillin-binding proteins. *J Comput Aided Mol Des.* **2014**, *29*(2), 127-141.
46. Nantasenamat C, Monnor T, Worachartcheewan A, Mandi P, Isarankura-Na-Ayudhya C, Prachayasittikul V. Predictive QSAR modeling of aldose reductase inhibitors using Monte Carlo feature selection. *Eur J Med Chem.* **2014**, *76*, 352-359.
47. Nantasenamat C, Worachartcheewan A, Mandi P, Monnor T, Isarankura-Na-Ayudhya C, Prachayasittikul V. QSAR modeling of aromatase inhibition by flavonoids using machine learning approaches. *Chemical Papers.* **2014**, *68*(5), 697-713.
48. Pingaew R, Mandi P, Nantasenamat C, Prachayasittikul S, Ruchirawat S, Prachayasittikul V. Design, synthesis and molecular docking studies of novel N-benzenesulfonyl-1,2,3,4-tetrahydroisoquinoline-based triazoles with potential anticancer activity. *Eur J Med Chem*. **2014**, *81*, 192-203.
49. Pingaew R, Saekee A, Mandi P, Nantasenamat C, Prachayasittikul S, Ruchirawat S, Prachayasittikul V. Synthesis, biological evaluation and molecular docking of novel chalcone–coumarin hybrids as anticancer and antimalarial agents. *Eur J Med Chem*. **2014**, *85*, 65-76.
50. Varnek A, Baskin I. Machine Learning Methods for Property Prediction in Chemoinformatics: Quo Vadis J Chem *Inf Model*. **2012**, *52*(6), 1413-1437.
51. Meredig B, Agrawal A, Kirklin S, Saal J, Doak J, Thompson A, Zhang K, Choudhary A, Wolverton C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B*. **2014**, *89*(9), 094104.
52. Isayev O, Fourches D, Muratov E, Oses C, Rasch K, Tropsha A, Curtarolo S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials*. **2015**, *27*(3), 735-743.

53. Oliynyk A, Antono E, Sparks T, Ghadbeigi L, Gaultois M, Meredig B, Mar A. High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chemistry of Materials*. **2016**, *28*(20), 7324-7331.
54. Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques M. Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chemistry of Materials*. **2017**, *29*(12), 5090-5103.
55. Pilania G, Mannodi-Kanakkithodi A, Uberuaga B, Ramprasad R, Gubernatis J, Lookman T. Machine learning bandgaps of double perovskites. *Sci Rep*. **2016**, *6*(1), 1-10.
56. Lee J, Seko A, Shitara K, Nakayama K, Tanaka I. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B.* **2016**, *93*(11), 115104.
57. Polishchuk P. Interpretation of Quantitative Structure–Activity Relationship Models: Past, Present, and Future. *J Chem Inf Model.* **2017**, *57*(11), 2618-2639.
58. Shoombuatong W, Prathipati P, Owasirikul W, Worachartcheewan A, Simeon S, Anuwongcharoen N, Wikberg J, Nantasenamat C. Towards the Revival of Interpretable QSAR Models. *Challenges and Advances in Computational Chemistry and Physics*. **2017**, *24*, 3-55.
59. Hendriks E, Kontogeorgis G, Dohrn R, de Hemptinne J, Economou I, Žilnik L, Vesovic V. Industrial Requirements for Thermodynamics and Transport Properties. *Ind Eng Chem Res*. **2010**, *49*(22), 11131-11141.
60. Meier R. A Way towards Reliable Predictive Methods for the Prediction of Physicochemical Properties of Chemicals Using the Group Contribution and other Methods. *Applied Sciences*. **2019**, *9*(8), 1700.
61. Fourches D, Muratov E, Tropsha A. Trust, But Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J Chem Inf Model*. **2010**, *50*(7), 1189-1204.
62. Doweyko A. QSAR: dead or alive? *J Comput Aided Mol Des*. **2008**, *22*(2), 81-89.
63. Dearden J, Cronin M, Kaiser K. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *SAR QSAR Environ Res*. **2009**, *20*(3-4), 241-266.
64. Le T, Epa V, Burden F, Winkler D. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem Rev.* **2012**, *112*(5), 2889-2919.
65. Frade A, McCabe P, Back K, Cooper R. Development of machine & deep learning models for the prediction of compounds' hydrogen-bond network dimensionality from their 2D molecular representations. Part II thesis, 2018. Unpublished.
66. Brennan, R. Generalizability theory. *Educational Measurement: Issues and Practice.* **1992**, *11*(4), 27-34.
67. Novotarskyi S, Sushko I, Körner R, Pandey A, Tetko I. A comparison of different QSAR approaches to modeling CYP450 1A2 inhibition. *J Chem Inf Model*. **2011**, *51*(6), 1271-1280.
68. Zhu H, Tropsha A, Fourches D, Varnek A, Papa E, Gramatica P, Öberg T, Dao P, Cherkasov A, Tetko I. Combinatorial QSAR Modeling of Chemical Toxicants Tested against Tetrahymena pyriformis. *J Chem Inf Model*. **2008**, *48*(4), 766-784.
69. Tropsha A, Gramatica P, Gombar V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb Sci*. **2003**, *22*(1), 69-77.
70. Tropsha A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol Inform.* **2010**, *29*(6-7), 476-488.
71. Dimova D, Bajorath J. Advances in Activity Cliff Research. *Mol Inform*. **2016**, *35*(5), 181-191.

72. Spjuth O, Willighagen E, Guha R, Eklund M, Wikberg J. Towards interoperable and reproducible QSAR analyses: Exchange of datasets. *J Cheminform*. **2010**, *2*(1), 1-7.
73. Worth A, Cronin M. Report of the Workshop on the Validation of QSARs and Other Computational Prediction Models. *Alternatives to Laboratory Animals.* **2004**, *32*, 703-706.
74. Nantasenamat C, Prachayasittikul V. Maximizing computational tools for successful drug discovery. *Expert Opin Drug Discov.* **2015**, *10*(4), 321-329.
75. Sushko I, Novotarskyi S, Körner R et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set. *J Chem Inf Model*. **2010**, *50*(12), 2094-2111.
76. Young D, Martin T, Venkatapathy R, Harten P. Are the Chemical Structures in Your QSAR Correct? *QSAR Comb Sci.* **2008**, *27*(11-12), 1337-1345.
77. Sheridan R. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J Chem Inf Model.* **2013**, *53*(4), 783-790.
78. Hawkins D. The Problem of Overfitting. *J Chem Inf Comput Sci*. **2003**, *44*(1), 1-12.
79. Mathea M, Klingspohn W, Baumann K. Chemoinformatic Classification Methods and their Applicability Domain. *Mol Inform.* **2016**, *35*(5), 160-180.
80. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research.* **2004**, *5*, 1205–1224.
81. Hanser T, Barber C, Marchaland J, Werner S. Applicability domain: towards a more formal definition. *SAR QSAR Environ Res.* **2016**, *27*(11), 865-881.
82. Sutton C, Boley M, Ghiringhelli L, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. *Nat Commun*. **2020**, *11*, 4428.
83. Sutton C, Ghiringhelli L, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, Golebiowski J, Liu X, Ziletti A, Scheffler M. Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition. *npj Computational Materials*. **2019**, *5*, 111.
84. Bryant M, Maloney A, Sykes R. Predicting mechanical properties of crystalline materials through topological analysis. *CrystEngComm*. **2018**, *20*(19), 2698-2704.
85. Shariare M, Leusen F, de Matas M, York P, Anwar J. Prediction of the Mechanical Behaviour of Crystalline Solids. *Pharm Res.* **2011**, *29*(1), 319-331.
86. Payne R, Roberts R, Rowe R, McPartlin M, Bashal A. The mechanical properties of two forms of primidone predicted from their crystal structures. *Int J Pharm*. **1996**, *145*(1-2), 165-173.
87. Chattoraj S, Shi L, Sun C. Understanding the relationship between crystal structure, plasticity and compaction behaviour of theophylline, methyl gallate, and their 1 : 1 co-crystal. *CrystEngComm*. **2010**, *12*(8), 2466-2472.
88. Bryant M, Rosbottom I, Bruno I, Docherty R, Edge C, Hammond R, Peeling R, Pickering J, Roberts K, Maloney A. “Particle Informatics”: Advancing Our Understanding of Particle Properties through Digital Design. *Cryst Growth Des*. **2019**, *19*(9), 5258-5266.
89. Reddy C, Rama Krishna G, Ghosh S. Mechanical properties of molecular crystals—applications to crystal engineering. *CrystEngComm.* **2010**, *12*(8), 2296-2314.
90. Sun C. Decoding Powder Tabletability: Roles of Particle Adhesion and Plasticity. *J Adhes Sci Technol*. **2011**, *25*(4-5), 483-499.
91. Reilly A, Cooper R, Adjiman C et al. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials*. **2016**, *72*(4), 439-459.
92. Dandekar P, Kuvadia Z, Doherty M. Engineering Crystal Morphology. Annu Rev Mater Res. **2013**, *43*, 359-386.

93. Variankaval N, Cote A, Doherty M. From form to function: Crystallization of active pharmaceutical ingredients. *AIChE Journal*. **2008**, *54*(7), 1682-1688.
94. Roberts K, Docherty R, Tamura R. Engineering Crystallography: From Molecule to Crystal To Functional Form. Springer, The Netherlands. **2017**, 109-131.
95. Friedel G. Études sur la loi de Bravais. *Bulletin de la Société française de Minéralogie.* **1907**, *30*(9), 326-455.
96. Bravais A. Études Cristallographiques. Gauthier-Villars, Paris. **1866**, 256-303.
97. Grimbergen R, Meekes H, Bennema P, Strom C, Vogels L. On the Prediction of Crystal Morphology. I. The Hartman–Perdok Theory Revisited. *Acta Crystallographica Section A Foundations of Crystallography.* **1998**, *54*(4), 491-500.
98. Censi R, Di Martino P. Polymorph Impact on the Bioavailability and Stability of Poorly Soluble Drugs. *Molecules.* **2015**, *20*(10), 18759-18776.
99. Lee E. A practical guide to pharmaceutical polymorph screening & selection. *Asian J Pharm Sci*. **2014**, *9*(4), 163-175.
100. Price S. The computational prediction of pharmaceutical crystal structures and polymorphism. *Adv Drug Deliv Rev*. **2004**, *56*(3), 301-319.
101. Beran G. Modeling Polymorphic Molecular Crystals with Electronic Structure Theory. *Chem Rev.* **2016**, *116*(9), 5567-5613.
102. Irwin J, Shoichet B. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J Chem Inf Model.* **2005**, *45*(1), 177-182.
103. Groom C, Bruno I, Lightfoot M, Ward S. The Cambridge Structural Database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials.* **2016**, *72*(2), 171-179.
104. Frade A, McCabe P, Cooper R. Increasing the performance, trustworthiness and practical value of machine learning models: a case study predicting hydrogen bond network dimensionalities from molecular diagrams. *CrystEngComm*. **2020**, *22*(43), 7186-7192.
105. Yousef M, Jung S, Showe L, Showe M. Learning from positive examples when the negative class is undetermined- microRNA gene identification. *Algorithms for Molecular Biology*. **2008**, *3*(1), 1-9.
106. Willett P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J Med Chem*. **2005**, *48*(13), 4183-4199.
107. Schaller D, Šribar D, Noonan T, Deng L, Nguyen T, Pach S, Machalz D, Bermudez M, Wolber G. Next generation 3D pharmacophore modeling. *WIREs Computational Molecular Science.* **2020**, *10*(4), e1468.
108. Raymond J, Willett P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J Comput Aided Mol Des*. **2002**, *16*(1), 59-71.
109. Rogers D, Hahn M. Extended-Connectivity Fingerprints. *J Chem Inf Model*. **2010**, *50*(5), 742-754.
110. Todeschini R, Consonni V. Handbook Of Molecular Descriptors. Wiley-VCH, Weinheim. **2000**, 124-226.
111. Danishuddin, Khan A. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. *Drug Discov Today.* **2016**, *21*(8), 1291-1302.
112. Gozalbes R, Doucet J, Derouin F. Application of Topological Descriptors in QSAR and Drug Design: History and New Trends. *Current Drug Target-Infectious Disorders*. **2002**, *2*(1), 93-102.
113. Sheridan R. Why do we need so many chemical similarity search methods? *Drug Discov Today*. **2002**, *7*(17), 903-911.
114. Dastmalchi S, Hamzeh-Mivehroud M, Asadpour-Zeynali K. Comparison of Different 2D and 3D-QSAR Methods on Activity Prediction of Histamine H3 Receptor Antagonists. *Iran J Pharm Res*. **2012**, *11*(1), 97-108.

115. Oprea T. On the information content of 2D and 3D descriptors for QSAR. *J Braz Chem Soc*. **2002**, *13*(6), 811-815.
116. Andrade C, Pasqualoto K, Ferreira E, Hopfinger A. 4D-QSAR: Perspectives in Drug Design. *Molecules.* **2010**, *15*(5), 3281-3294.
117. Sun C, Kiang Y. On the Identification of Slip Planes in Organic Crystals Based on Attachment Energy Calculation. *J Pharm Sci.* **2008**, *97*(8), 3456-3461.
118. Venkatraman V, Pérez-Nueno V, Mavridis L, Ritchie D. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *J Chem Inf Model.* **2010**, *50*(12), 2079-2093.
119. Brown R, Martin Y. Use of Structure−Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *J Chem Inf Comput Sci*. **1996**, *36*(3), 572-584.
120. Matter H, Pötter T. Comparing 3D Pharmacophore Triplets and 2D Fingerprints for Selecting Diverse Compound Subsets. *J Chem Inf Comput Sci*. **1999**, *39*(6), 1211-1225.
121. Estrada E, Molina E, Perdomo-López I. Can 3D Structural Parameters Be Predicted from 2D (Topological) Molecular Descriptors? *J Chem Inf Comput Sci*. **2001**, *41*(4), 1015-1021.
122. Xue L, Stahura F, Godden J, Bajorath J. Mini-fingerprints Detect Similar Activity of Receptor Ligands Previously Recognized Only by Three-Dimensional Pharmacophore-Based Methods. *J Chem Inf Comput Sci*. **2001**, *41*(2), 394-401.
123. Raccuglia P, Elbert K, Adler P, Falk C, Wenny M, Mollo A, Zeller M, Friedler S, Schrier J, Norquist A. Machine-learning-assisted materials discovery using failed experiments. *Nature.* **2016**, *533*(7601), 73-76.
124. Williams C. Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. *J Am Stat Assoc.* **2003**, *98*(462), 489-489.
125. Smola A, Schölkopf B. A tutorial on support vector regression. *Stat Comput*. **2004**, *14*(3), 199-222.
126. Burges C. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov*. **1998**, *2*(2), 121-167.
127. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manag.* **2009**, *45*(4), 427-437.
128. Zakharov A, Peach M, Sitzmann M, Nicklaus M. QSAR Modeling of Imbalanced High-Throughput Screening Data in PubChem. *J Chem Inf Model.* **2014**, *54*(3), 705-712.
129. Roberts K, Sherwood J, Yoon C, Docherty R. Understanding the Solvent-Induced Habit Modification of Benzophenone in Terms of Molecular Recognition at the Crystal/Solution Interface. *Chemistry of Materials.* **1994**, *6*(8), 1099-1102.
130. Bennema P. Crystal growth from solution - Theory and experiment. *J Cryst Growth*. **1974**, *24-25*, 76-83.
131. Cruz-Cabeza A, Reutzel-Edens S, Bernstein J. Facts and fictions about polymorphism. *Chem Soc Rev.* **2015**, *44*(23), 8619-8635.
132. van der Maaten L, Hinton G. Visualizing Data using t-SNE. *Journal of Machine Learning Research*. **2008**, *1*, 1-48.
133. RDKit: Open-source cheminformatics. Available at: http://www.rdkit.org. (last accessed 21 Sep 2021)
134. Fourches D, Muratov E, Tropsha A. Curation of chemogenomics data. *Nat Chem Biol*. **2015**, *11*(8), 535-535.
135. Fourches D, Muratov E, Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J Chem Inf Model*. **2016**, *56*(7), 1243-1252.

136. Macrae C, Bruno I, Chisholm J, Edgington P, McCabe P, Pidcock E, Rodriguez-Monge L, Taylor R, van de Streek J, Wood P. Mercury CSD 2.0– new features for the visualization and investigation of crystal structures. *J Appl Crystallogr*. **2008**, *41*(2), 466-470.
137. Rücker C, Rücker G, Meringer M. y-Randomization and Its Variants in QSPR/QSAR. *J Chem Inf Model.* **2007**, *47*(6), 2345-2357.
138. Montáns F, Chinesta F, Gómez-Bombarelli R, Kutz J. Data-driven modeling and learning in science and engineering. *Comptes Rendus Mécanique*. **2019**, *347*(11), 845-855.
139. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Wiley-VCH, Weinheim. **2009**, 39-75.
140. Mauri A, Consonni V, Pavan M, Todeschini R. DRAGON software: An easy approach to molecular descriptor calculations. *Commun Math Comput Chem*. **2006**, *56*, 237-248.
141. Labute P. A widely applicable set of descriptors. *Journal of Molecular Graphics and Modelling*. **2000**, *18*(4-5), 464-477.
142. Hall L, Mohney B, Kier L. The Electrotopological State: An Atom Index for QSAR. *Quantitative Structure-Activity Relationships.* **1991**, *10*(1), 43-51.
143. Gasteiger J, Marsili M. A new model for calculating atomic charges in molecules. *Tetrahedron Lett.* **1978**, *19*(34), 3181-3184.
144. Whitley D. Van der Waals surface graphs and molecular shape. *J Math Chem*. **1998**, *23*(3-4), 377-397.
145. Bonchev D, Trinajstić N. Information theory, distance matrix, and molecular branching. *J Chem Phys*. **1977**, *67*(10), 4517-4533.
146. Ertl P, Rohde B, Selzer P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J Med Chem*. **2000**, *43*(20), 3714-3717.
147. Wildman S, Crippen G. Prediction of Physicochemical Parameters by Atomic Contributions. *J Chem Inf Comput Sci.* **1999**, *39*(5), 868-873.
148. Abraham M, Platts J. Hydrogen Bond Structural Group Constants. *J Org Chem*. **2001**, *66*(10), 3484-3491.
149. Hunter C. Quantifying Intermolecular Interactions: Guidelines for the Molecular Recognition Toolbox. *Angewandte Chemie International Edition*. **2004**, *43*(40), 5310-5324.
150. Jencks W. On the attribution and additivity of binding energies*. Proceedings of the National Academy of Sciences.* **1981**, *78*(7), 4046-4050.
151. Schneider H. Linear free energy relationships and pairwise interactions in supramolecular chemistry. *Chem Soc Rev.* **1994**, *23*(4), 227-234.
152. H. Williams D, S. Westwell M. Aspects of weak interactions. *Chem Soc Rev*. **1998**, *27*, 57-64.
153. Hirschfelder J. Advances in Chemical Physics: Intermolecular Forces. Wiley, New York NY. **1967**, 643-645
154. Bisson A, Hunter C, Morales J, Young K. Cooperative Interactions in a Ternary Mixture. *Chemistry – A European Journal.* **1998**, *4*(5), 845-851.
155. Chapman K, Still W. A remarkable effect of solvent size on the stability of a molecular complex. *J Am Chem Soc.* **1989**, *111*(8), 3075-3077.
156. Venkatesh B, Anuradha J. A Review of Feature Selection and Its Methods. *Cybernetics and Information Technologies.* **2019**, *19*(1), 3-26.
157. Hall M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In: Proceedings of the 17th international conference on machine learning. **2000**, 359-366.
158. Kohavi R, John G. Wrappers for feature subset selection. *Artif Intell*. **1997**, *97*(1-2), 273-324.

159. Ferguson L. Hydrogen bonding and physical properties of substances. *J Chem Educ*. **1956**, *33*(6), 267.
160. Yin P, Fan X. Estimating R2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods. *The Journal of Experimental Education.* **2010**, *69*(2), 203-224.
161. Glantz S, Slinker B. Primer Of Applied Regression & Analysis Of Variance. McGraw-Hill, New York, NY. **1990**, 57-90.
162. Li X, Sha J, Wang Z. A comparative study of multiple linear regression, artificial neural network and support vector machine for the prediction of dissolved oxygen. *Hydrology Research*. **2017**, *48*(5), 1214-1225.
163. Parveen N, Zaidi S, Danish M. Support vector regression model for predicting the sorption capacity of lead (II). *Perspect Sci.* **2016**, *8*, 629-631.
164. Azamathulla H, Wu F. Support vector machine approach for longitudinal dispersion coefficients in natural streams. *Appl Soft Comput*. **2011**, *11*(2), 2902-2905.
165. Dietterich T. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*(7), 1895-1923.
166. Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*. **1945**, *1*(6), 80-83.
167. Demsar J. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*. **2006**, *7*, 1–30.
168. Lipinski C, Hopkins A. Navigating chemical space for biology and medicine. *Nature*. **2004**, *432*, 855-861.
169. Reymond J, van Deursen R, Blum L, Ruddigkeit L. Chemical space as a source for new drugs. *Medchemcomm.* **2010**, *1*, 30-38.
170. Goodnow R, Dumelin C, Keefe A. DNA-encoded chemistry: enabling the deeper sampling of chemical space. *Nature Reviews Drug Discovery*. **2017**, *16*, 131-147.
171. Reymond J, Ruddigkeit L, Blum L, van Deursen R. The enumeration of chemical space. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. **2012**, *2*(5), 717-733.
172. Bohacek R, McMartin C, Guida W. The art and practice of structure-based drug design: A molecular modeling perspective. *Med Res Rev*. **1996**, *16*(1), 3-50.
173. Polishchuk P, Madzhidov T, Varnek A. Estimation of the size of drug-like chemical space based on GDB-17 data. *J Comput Aided Mol Des*. **2013**, *27*(8), 675-679.
174. Fink T, Bruggesser H, Reymond J. Virtual Exploration of the Small-Molecule Chemical Universe below 160 Daltons. *Angewandte Chemie International Edition*. **2005**, *44*(10), 1504-1508.
175. Blum L, Reymond J. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J Am Chem Soc*. **2009**, *131*(25), 8732-8733.
176. Ruddigkeit L, van Deursen R, Blum L, Reymond J. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J Chem Inf Model*. **2012**, *52*(11), 2864-2875.
177. Awale M, Visini R, Probst D, Arús-Pous J, Reymond J. Chemical Space: Big Data Challenge for Molecular Diversity. *CHIMIA International Journal for Chemistry*. **2017**, *71*(10), 661-666.
178. Dobson C. Chemical space and biology. *Nature*. **2004**, *432*(7019), 824-828.
179. Bartlett P, Entzeroth M. Chemical diversity: Definition and quantification. In: Exploiting Chemical Diversity for Drug Discovery. Royal Society of Chemistry, Cambridge, UK. **2006**, 137–160.
180. Gong Z, Hu G, Li Q, Liu Z, Wang F, Zhang X, Xiong J, Li P, Xu Y, Ma R, Chen S, Li J. Compound Libraries: Recent Advances and Their Applications in Drug Discovery. *Curr Drug Discov Technol.* **2017**, *14*(4), 216-228.
181. Cone M, Venkataraghavan R, McLafferty F. Molecular structure comparison program for the identification of maximal common substructures. *J Am Chem Soc*. **1977**, *99*(23), 7668-7671.

182. Wang J, Ge Y, Xie X. Development and Testing of Druglike Screening Libraries. *J Chem Inf Model*. **2019**, *59*(1), 53-65.
183. Cortes-Ciriano I. Benchmarking the Predictive Power of Ligand Efficiency Indices in QSAR. *J Chem Inf Model*. **2016**, *56*(8), 1576-1587.
184. Paricharak S, IJzerman A, Jenkins J, Bender A, Nigsch F. Data-Driven Derivation of an "Informer Compound Set" for Improved Selection of Active Compounds in High-Throughput Screening. *J Chem Inf Model*. **2016**, *56*(9), 1622-1630.
185. Akella L, DeCaprio D. Cheminformatics approaches to analyze diversity in compound screening libraries. *Curr Opin Chem Biol.* **2010**, *14*(3), 325-330.
186. Bartlett P, Foster D, Telgarsky M. Spectrally-normalized margin bounds for neural networks. **2017**. *arXiv preprint arXiv: 1706.08498.*
187. Shawe-Taylor J, Rivasplata O. Statistical Learning Theory: A Hitchhiker's Guide. *Neural Inf Process Syst.* **2018**, *12*, 1-51.
188. Reid M. Generalization Bounds. In: Encyclopedia of Machine Learning. Springer, Boston, MA. **2011**
189. Bousquet, O, Boucheron, S, Lugosi, G. Introduction to Statistical Learning Theory. In: Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg. **2003**, 169-207
190. Zhang Y, Wu H, Liu H, Tong L, Wang M. Improve model generalization and robustness to dataset bias with bias-regularized learning and domain-guided augmentation. **2019**. *arXiv preprint arXiv:191006745*.
191. Jin W, Barzilay R, Jaakkola T. Domain extrapolation via regret minimization. **2020**. *arXiv preprint arXiv:200603908*.
192. Vapnik V, Chervonenkis A. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability & Its Applications*. **1971**, *16*(2), 264-280.
193. Bartlett P, Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*. **2002**, *3*, 463-482.
194. McAllester D. Some PAC-Bayesian theorems. *Mach Learn*. **1999**, *37*(3), 355-363.
195. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. **2016**. *arXiv preprint arXiv:61103530*.
196. Rios J, Li X, Faulkner-Bond M. A review of methodological trends in generalizability theory. *Northeastern Educational Research Association Conference*. **2012**.
197. Glavatskikh M, Leguy J, Hunault G, Cauchy T, Da Mota B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J Cheminform*. **2019**, *11*, 69.
198. Eriksson L, Jaworska J, Worth A, Cronin M, McDowell R, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect.* **2003**, *111*(10), 1361-1375.
199. Dragos H, Gilles M, Alexandre V. Predicting the Predictability: A Unified Approach to the Applicability Domain Problem of QSAR Models. *J Chem Inf Model*. **2009**, *49*(7), 1762-1776.
200. Sheridan R. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J Chem Inf Model.* **2012**, *52*(3), 814-823.
201. Tetko I, Bruneau P, Mewes H, Rohrer D, Poda G. Can we estimate the accuracy of ADME–Tox predictions? *Drug Discov Today*. **2006**, *11*(15-16), 700-707.
202. Sahigara F, Mansouri K, Ballabio D, Mauri A, Consonni V, Todeschini R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules*. **2012**, *17*(5), 4791-4810.

203. Toplak M, Močnik R, Polajnar M, Bosnić Z, Carlsson L, Hasselgren C, Demšar J, Boyer S, Zupan B, Stålring J. Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. *J Chem Inf Model.* **2014**, *54*(2), 431-441.
204. Carrió P, Pinto M, Ecker G, Sanz F, Pastor M. Applicability Domain Analysis (ADAN): A Robust Method for Assessing the Reliability of Drug Property Predictions. *J Chem Inf Model*. **2014**, *54*(5), 1500-1511.
205. Sheridan R. The Relative Importance of Domain Applicability Metrics for Estimating Prediction Errors in QSAR Varies with Training Set Diversity. *J Chem Inf Model*. **2015**, *55*(6), 1098-1107.
206. Roy K, Kar S, Ambure P. On a simple approach for determining applicability domain of QSAR models. *Chemometrics and Intelligent Laboratory Systems.* **2015**, *145*, 22-29.
207. Norinder U, Rybacka A, Andersson P. Conformal prediction to define applicability domain – A case study on predicting ER and AR binding. *SAR QSAR Environ Res*. **2016**, *27*(4), 303-316.
208. Kar S, Roy K, Leszczynski J. Applicability Domain: A Step Toward Confident Predictions and Decidability for QSAR Modeling. In: Computational Toxicology. Methods in Molecular Biology. Humana Press, New York, NY. **2018**, 141-169.
209. Roy K, Kar S, Das R. Understanding The Basics Of QSAR For Applications In Pharmaceutical Sciences And Risk Assessment. Cambridge, Massachusetts: Academic Press; 2015.
210. Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR Applicability Domain Estimation by Projection of the Training Set in Descriptor Space: A Review. *Alternatives to Laboratory Animals*. **2005**, *33*(5), 445-459.
211. Netzeva T, Worth A, Aldenberg T, Benigni R, Cronin M, Gramatica P, Jaworska J, Kahn S, Klopman G, Marchant C, Myatt G. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Alternatives to Laboratory Animals*. **2005**, *33*(2), 155-173.
212. Tetko I, Sushko I, Pandey A, Zhu H, Tropsha A, Papa E, Öberg T, Todeschini R, Fourches D, Varnek A. Critical Assessment of QSAR Models of Environmental Toxicity against Tetrahymena pyriformis: Focusing on Applicability Domain and Overfitting by Variable Selection. *J Chem Inf Model.* **2008**, *48*(9), 1733-1746.
213. Klingspohn W, Mathea M, ter Laak A, Heinrich N, Baumann K. Efficiency of different measures for defining the applicability domain of classification models. *J Cheminform*. **2017**, *9*, 44.
214. Fechner N, Jahn A, Hinselmann G, Zell A. Estimation of the applicability domain of kernel-based machine learning models for virtual screening. *J Cheminform*. **2010**, *2*, 2.
215. Sushko I, Novotarskyi S, Körner R, Pandey A, Kovalishyn V, Prokopenko V, Tetko I. Applicability domain for in silico models to achieve accuracy of experimental measurements. *J Chemom*. **2010**, *24*(3-4), 202-208.
216. Hodge V, Austin J. A Survey of Outlier Detection Methodologies. *Artif Intell Rev*. **2004**, *22*(2), 85-126.
217. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv.* **2009**, *41*(3), 1-58.
218. Pimentel M, Clifton D, Clifton L, Tarassenko L. A review of novelty detection. *Signal Processing*. **2014**, *99*, 215-249.
219. Nouretdinov I, Li G, Gammerman A, Luo Z. Application of Conformal Predictors to Tea Classification Based on Electronic Nose. In: Artificial Intelligence Applications and Innovations. AIAI 2010. IFIP Advances in Information and Communication Technology. Springer, Berlin, Heidelberg. **2010**, 303-310.
220. Liu Y. A Comparative Study on Feature Selection Methods for Drug Discovery. *J Chem Inf Comput Sci*. **2004**, *44*(5), 1823-1828.

221. Vega J, Murari A, Pereira A, González S, Pastor I. Accurate and reliable image classification by using conformal predictors in the TJ-II Thomson scattering. *Review of Scientific Instruments*. **2010**, *81*(10), 10E118.
222. Shafer G, Vovk V. A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*. **2008**, *9*(12), 371-421.
223. Ivina O, Nouretdinov I, Gammerman A. Valid predictions with confidence estimation in an air pollution problem. *Progress in Artificial Intelligence*. **2012**, *1*(3), 235-243.
224. AD Software. Available at: http://teqip.jdvu.ac.in/QSAR_Tools/. (last accessed 6 Sep 2021)
225. QSARINS. Available at: http://www.qsar.it/. (last accessed 6 Sep 2021)
226. Domain Manager. Available at: http://oasis-lmc.org/products/software/domain-manager.aspx. (last accessed 6 Sep 2021)
227. Atkinson A, Hawkins D. Identification of Outliers. *Biometrics*. **1981**, *37*(4), 860-865.
228. Maletic J, Marcus A. Data Cleansing: Beyond Integrity Analysis. *Iq*. **2000**, 200-209.
229. Zimek A, Schubert E, Kriegel H. A survey on unsupervised outlier detection in high-dimensional numerical data. *Stat Anal Data Min*. **2012**, *5*(5), 363-387.
230. Escalante H. A comparison of outlier detection algorithms for machine learning. In: Proceedings of the 2nd International Conference on Deep Learning Technologies. **2018**, 228-237.
231. Nazari Z, Yu S, Kang D, Kawachi Y. Comparative Study of Outlier Detection Algorithms for Machine Learning. In: Proceedings of the 2nd International Conference on Deep Learning Technologies. **2018**, 47-51.
232. Kolbaşi A, Ünsal, A. A comparison of the outlier detecting methods: an application on Turkish foreign trade data. *Journal of Mathematics and Statistical Science*. **2015**, *5*, 213-234.
233. Clarke R, Ressom H, Wang A, Xuan J, Liu M, Gehan E, Wang Y. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*. **2008**, *8*(1), 37-49.
234. Menard P, Mason J, Morize I, Bauerschmidt S. Chemistry Space Metrics in Diversity Analysis, Library Design, and Compound Selection. *J Chem Inf Comput Sci*. **1998**, *38*(6), 1204-1213.
235. Woźniak M, Wołos A, Modrzyk U, Górski R, Winkowski J, Bajczyk M, Szymkuć S, Grzybowski B, Eder M. Linguistic measures of chemical diversity and the "keywords" of molecular collections. *Sci Rep*. **2018**, *8*, 7598.
236. Koutsoukas A, Paricharak S, Galloway W, Spring D, IJzerman A, Glen R, Marcus D, Bender A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J Chem Inf Model*. **2014**, *54*(1), 230-242.
237. González-Medina M, Prieto-Martínez F, Owen J, Medina-Franco J. Consensus Diversity Plots: a global diversity analysis of chemical libraries. *J Cheminform*. **2016**, *8*, 63.
238. Liu F, Ting K, Zhou Z. Isolation Forest. In: Proceedings of the 8th IEEE International Conference on Data Mining. **2008**, 413-422.
239. Liu F, Ting K, Zhou Z. Isolation-Based Anomaly Detection. *ACM Trans Knowl Discov Data*. **2012**, *6*(1), 1-39.
240. Breunig M, Kriegel H, Ng R, Sander J. LOF: Identifying density-based local outliers. *ACM SIGMOD Record*. **2000**, *29*(2), 93-104.
241. Schölkopf B, Platt J, Shawe-Taylor J, Smola A, Williamson R. Estimating the Support of a High-Dimensional Distribution. *Neural Comput*. **2001**, *13*(7), 1443-1471.
242. Kriegel H, Schubert M, Zimek A. Angle-based outlier detection in high-dimensional data. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **2008**, 444–452.

243. Goldstein M, Dengel A. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track, 9*. **2012**.
244. Hadi A. Identifying Multiple Outliers in Multivariate Data. *Journal of the Royal Statistical Society: Series B (Methodological).* **1992**, *54*(3), 761-771.
245. Platt J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*. **1999**, *10*(3), 61-74.
246. Menon A, Jiang X, Vembu S, Elkan C, Ohno-Machado L. Predicting accurate probabilities with a ranking loss. In: Proceedings of the 29th International Conference on Machine Learning. **2012**, 703.
247. Kull M, Silva Filho T, Flach P. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electron J Stat.* **2017**, *11*(2), 5052-5080.
248. Wu T, Lin C, Weng R. Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research.* **2004**, *5*, 975–1005.
249. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **2002**, 694–699.
250. Guyon I, Elisseeff, A, Kaelbling L. An introduction to variable and feature selection. *Journal of Machine Learning Research.* **2003**, *3*(7-8), 1157–1182.
251. Smialowski P, Frishman D, Kramer S. Pitfalls of supervised feature selection. *Bioinformatics.* **2009**, *26*(3), 440-443.
252. Liu C, Jiang D, Yang W. Global geometric similarity scheme for feature selection in fault diagnosis. *Expert Syst Appl.* **2014**, *41*(8), 3585-3595.
253. Abusamra H. A Comparative Study of Feature Selection and Classification Methods for Gene Expression Data of Glioma. *Procedia Comput Sci*. **2013**, *23*, 5-14.
254. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol.* **2005**, *3*(2), 185-205.
255. Miao J, Niu L. A Survey on Feature Selection. *Procedia Comput Sci*. **2016**, *91*, 919-926.
256. Jovic A, Brkic K, Bogunovic N. A review of feature selection methods with applications. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*. **2015**, 1200-1205.
257. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics.* **2007**, *23*(19), 2507-2517.
258. Anukrishna P, Paul V. A review on feature selection for high dimensional data. *2017 International Conference on Inventive Systems and Control.* **2017**, 1-4.
259. Guyon I, Elisseeff A. An Introduction to Feature Extraction. In: Feature Extraction. Studies in Fuzziness and Soft Computing. Springer, Berlin, Heidelberg. **2006**, 1-25.
260. Yu L, Liu H. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning. **2003**, 856-863.
261. Zare H, Haffari G, Gupta A, Brinkman R. Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis. BMC Genomics. **2013**, *14*, S14.
262. Alsenan S, Al-Turaiki I, Hafez A. Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study. *IEEE Access*. **2020**, *8*, 78737-78752.
263. Auffarth B, López M, Cerquides J. Comparison of Redundancy and Relevance Measures for Feature Selection in Tissue Classification of CT Images. In: Advances in Data Mining. Applications and Theoretical Aspects. Springer, Berlin, Heidelberg. **2010**, 248-262.
264. Chou T, Yen K, Luo J, Pissinou N, Makki K. Correlation-Based Feature Selection for Intrusion Detection Design. *MILCOM 2007 - IEEE Military Communications Conference*. **2007**, 1-7.

265. Czarnecki W, Jastrzebski S, Sieradzki I, Podlewska S. Active Learning of Compounds Activity–Towards Scientifically Sound Simulation of Drug Candidates Identification. In: Proceedings of MLLS: 2nd Workshop on Machine Learning in Life Sciences. **2015**, 40-51.
266. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de-novo design through deep reinforcement learning. *J Cheminform*. **2017**, *9*(1), 1-14.
267. Kuz'min V, Artemenko A, Muratov E, Volineckaya I, Makarov V, Riabova O, Wutzler P, Schmidtke M. Quantitative Structure−Activity Relationship Studies of [(Biphenyloxy)propyl]isoxazole Derivatives. Inhibitors of Human Rhinovirus 2 Replication. *J Med Chem.* **2007**, *50*(17), 4205-4213.
268. Lowe R, Mussa H, Nigsch F, Glen R, Mitchell J. Predicting the mechanism of phospholipidosis. *J Cheminform.* **2012**, *4*(1), 1-9.
269. Ploemen J, Kelder J, Hafmans T, van de Sandt H, van Burgsteden J, Salemink P, van Esch E. Use of physicochemical calculation of pKa and CLogP to predict phospholipidosis-inducing potential. *Experimental and Toxicologic Pathology*. **2004**, *55*(5), 347-355.
270. Carlsson L, Helgee E, Boyer S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J Chem Inf Model.* **2009**, *49*(11), 2551-2558.
271. Chen H, Carlsson L, Eriksson M, Varkonyi P, Norinder U, Nilsson I. Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms. *J Chem Inf Model*. **2013**, *53*(6), 1324-1336.
272. Muratov E, Artemenko A, Varlamova E, Polischuk P, Lozitsky V, Fedchuk A, Lozitska R, Gridina T, Koroleva L, Sil'nikov V, Galabov A, Makarov V, Riabova O, Wutzler P, Schmidtke M, Kuz'min V. Per aspera ad astra: application of Simplex QSAR approach in antiviral research. *Future Med Chem.* **2010**, *2*(7), 1205-1226.
273. Bhhatarai B, Garg R, Gramatica P. Are Mechanistic and Statistical QSAR Approaches Really Different? MLR Studies on 158 Cycloalkyl-Pyranones. *Mol Inform.* **2010**, *29*(6-7), 511-522.
274. Kuz'min V, Polishchuk P, Artemenko A, Andronati S. Interpretation of QSAR Models Based on Random Forest Methods. *Mol Inform.* **2011**, *30*(6-7), 593-603.
275. Garson G. Interpreting neural-network connection weights. *AI Expert.* **1991**, *6*(4), 46-51.
276. Goh A. Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering.* **1995**, *9*(3), 143-151.
277. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. **2006**, *7*(1), 1-13.
278. So S, Richards W. Application of neural networks: quantitative structure-activity relationships of the derivatives of 2,4-diamino-5-(substituted-benzyl)pyrimidines as DHFR inhibitors. *J Med Chem*. **1992**, *35*(17), 3201-3207.
279. Aoyama T, Ichikawa H. Neural networks as nonlinear structure-activity relationship analyzers. Useful functions of the partial derivative method in multilayer neural networks. *J Chem Inf Model*. **1992**, *32*(5), 492-500.
280. Lundberg S, Lee S. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. **2017**, 4768–4777.
281. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **2016**, 13-17.
282. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*. **2005**, *27*(8), 1226-1238.
283. Kvalseth T. Entropy and Correlation: Some Comments. *IEEE Trans Syst Man Cybern*. **1987**, *17*(3), 517-519.

284. Vinh N, Epps J, Bailey J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*. **2010**, *11*(95), 2837−2854.
285. Vinh N, Chan J, Bailey J. Reconsidering mutual information based feature selection: A statistical significance view. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence. **2014**, 1.
286. Nguyen X, Chan J, Romano S, Bailey J. Effective global approaches for mutual information based feature selection. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. **2014**, 512–521.
287. Estevez P, Tesmer M, Perez C, Zurada J. Normalized Mutual Information Feature Selection. *IEEE Trans Neural Netw.* **2009**, *20*(2), 189-201.
288. Garcia S, Luengo J, Sáez J, López V, Herrera F. A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Trans Knowl Data Eng*. **2013**, *25*(4), 734-750.
289. Jirapech-Umpai T, Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*. **2005**, *6*, 148.
290. Xuan P, Guo M, Wang J, Wang C, Liu X, Liu Y. Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. *Genetics and Molecular Research*. **2011**, *10*(2), 588-603.
291. Peng S, Xu Q, Ling X, Peng X, Du W, Chen L. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Lett*. **2003**, *555*(2), 358-362.
292. Vermeulen-Jourdan L, Dhaenens C, Talbi E. Linkage disequilibrium study with a parallel adaptive GA. *International Journal of Foundations of Computer Science*. **2005**, *16*(2), 241-260.
293. Nason G, Scott D. Multivariate Density Estimation: Theory, Practice, and Visualization. *Journal of the Royal Statistical Society Series A (Statistics in Society).* **1993**, *156*(3), 511-512.
294. Yu L, Reutzel-Edens S, Mitchell C. Crystallization and Polymorphism of Conformationally Flexible Molecules: Problems, Patterns, and Strategies. *Org Process Res Dev*. **2000**, *4*(5), 396-402.
295. Sessler C, Rahm M, Becker S, Goldberg J, Wang F, Lippard S. CF2H, a Hydrogen Bond Donor. *J Am Chem Soc.* **2017**, *139*(27), 9325-9332.
296. Alabugin I, Manoharan M, Peabody S, Weinhold F. Electronic Basis of Improper Hydrogen Bonding: A Subtle Balance of Hyperconjugation and Rehybridization. *J Am Chem Soc*. **2003**, *125*(19), 5973-5987.
297. de la Iglesia B. Evolutionary computation for feature selection in classification problems. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.* **2013**, *3*(6), 381-407.
298. Lipinski C, Lombardo F, Dominy B, Feeney P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*. **1997**, *23*(1-3), 3-25.
299. Cameron A, Windmeijer F. An R-squared measure of goodness of fit for some common nonlinear regression models. J Econom. 1997;77(2):329-342. 1. Varnek A, Baskin I. Chemoinformatics as a Theoretical Chemistry Discipline. *Mol Inform*. **2011**, *30*(1), 20-32.

Supplementary Information

A. Molecular representations

The low-level representation describes molecules as 2048-bit long binary extended connectivity fingerprints of radius 3.[109] The mid-level representation describes molecules as a collection of 175 structural features, accounting for 170

functional group counts and the 5 Lipinski properties.[298] The high-level representation describes molecules as a collection of all 113 two and lower dimensional molecular descriptors from the RDKit library, including atom counts and topological indices. A full list of mid- and high- descriptors can be found below.

List of the high-level representation topological descriptors.

MolWt	RingCount	MinEStateIndex
ExactMolWt	NumAliphaticRings	PEOE_VSA1-14
HeavyAtomMolWt	NumAromaticRings	SMR_VSA1-10
NumHAcceptors	NumSaturatedRings	SlogP_VSA1-12
NumHDonors	NumValenceElectrons	EState_VSA1-11
NumRotatableBonds	NumRadicalElectrons	VSA_EState1-10
FractionCSP3	qed	LabuteASA
NHOHCount	MolLogP	Chi0-1
NOCount	MolMR	Chi0n-4n
NumHeteroatoms	BalabanJ	Chi0v-4v
HeavyAtomCount	MaxAbsPartialCharge	HallKierAlpha
NumAliphaticCarbocycles	MinAbsPartialCharge	Kappa1-3
NumAromaticCarbocycles	MaxPartialCharge	BertzCT
NumSaturatedCarbocycles	MinPartialCharge	TPSA
NumAliphaticHeterocycles	MaxAbsEStateIndex	Ipc
NumAromaticHeterocycles	MinAbsEStateIndex	FpDensityMorgan1-3
NumSaturatedHeterocycles	MaxEStateIndex	

List of the mid-level representation descriptors.

MolWt	Amide count	CycloheximideDeriv count
MolLogP	Amidine count	MonensinDeriv count
NumHAcceptors	Guanidine count	CyanidolDeriv count
NumHDonors	H-PyrroleN count	SqualestatinDeriv count
NumRotatableBonds	Imide count	Thionyl count
AcidChloride Aromatic count	Thiocyanate count	Anhydride count
AcidChloride Aliphatic count	Sulfide count	Mustard count
CarboxylicAcid Aromatic count	Sulfone count	HydrocarbonTails count
CarboxylicAcid Aliphatic count	Barbiturate count	Benzamidine count
SulfonylChloride Aromatic count	Urea count	Hydroxamate count
SulfonylChloride Aliphatic count	Imidazole count	Acetal count
Amine Primary Aromatic count	Furan count	Alkyne count
Amine Primary Aliphatic count	Thiophene count	AminoAcid count
Amine Secondary Aromatic count	Thiazole count	Dipeptide count
Amine Secondary Aliphatic count	Oxazole count	Enamine count
Amine Tertiary Aromatic count	Pyridine count	Enol count
Amine Tertiary Aliphatic count	Piperdine count	Hemiacetal count
BoronicAcid Aromatic count	Piperzine count	Hemiketal count

BoronicAcid Aliphatic count
Isocyanate Aromatic count
Isocyanate Aliphatic count
Alcohol Aromatic count
Alcohol Aliphatic count
Aldehyde Aromatic count
Aldehyde Aliphatic count
Halogen Aromatic count
Halogen Aliphatic count
Azide Aromatic count
Azide Aliphatic count
Nitro Aromatic count
Nitro Aliphatic count
Isothiocyanate count
Nitroso count
Oxime count
Hydrazine count
Diazo count
Cyano count
PrimarySulfonamide count
SulfonicAcid count
MethylSulfinyl count
MethylThio count
Thiol count
Thiocarbonyls count
MethylSulfonamide count
Tbutyl count
Trifluoromethyl count
Acetylene count
Cyclopropyl count
Ethoxy count
Methoxy count
Nitrile count
Carbonyl count
Ester count
Ketone count
Phenol count
Aniline count
Imine count
Hidrazone count
Hydroxylamine count
Morpholine count
Lactams count
Lactone count
Tetrazole count
Epoxide count
AlkaneUnbranched count
Bicyclic count
Benzene count
PhosphoricAcid count
PhosphoricEster count
Benzodiazepine count
NitroBenzeneSubst count
NitroAromaticSubst count
Dihydropyridines count
ParaHydroxylation count
AllylicOxSites count
XCCNR count
AcidHalide count
Carbazide count
SulphateEster count
Sulphonate count
Peroxide count
PentafluorophenylEster count
Triflate count
PhosphorAmide count
AcylHydrazide count
Phosphorane count
Chloramidine count
Triacyloxime count
Cyanohydrin count
Cyanophosphonate count
Azocyanamide count
Azoalkanal count
Disulphide count
Hydrazothiourea count
Thioester count
Cyanamide count
Quinone count
Polyenes count
SaponinDeriv count
CytochalasinDeriv count
MalonicAcid count
Macrocycle count
Oxalyl count
Thioamide count
Ketene count
Hemiaminal count
Aminal count
Thiohemiaminal count
Thioacetal count
Enediol count
EtherDialkyl count
EtherAlkylaryl count
EtherDiaryl count
ThioetherDialkyl count
ThioetherAlkylaryl count
ThioetherDiaryl count
Alkene count
Allene count
CarbothioicAcid count
CarbothioicSEster count
CarbothioicSLactone count
CarbothioicOEster count
CarbothioicOLactone count
CarbothioicHalide count
Amidrazone count
Urethan count
Biuret count
Carbazone count
Nitrite count
Nitrate count
Spiro count
AnnelatedRings count
BridgedRings count
SugarPattern1 count
SugarPattern2 count
Conjugated_double_bond count
Conjugated_tripple_bond count
Cis_double_bond count
Trans_double_bond count

B. Estimator and hyperparameter optimisation results

Statistical methods considered for model selection. Estimators are sorted in descending order of accuracy, hyperparameters shown where grid search optimised, whilst those not shown were set to scikit-learn default values.

HBND model framework study results.

Estimator	Hyperparameters	Accuracy	F1 score
SVM Classifier	kernel = linear, C = 1, gamma = 0.01	0.55	0.55
Random Forest Classifier	max_depth = 50, n_estimators = 100	0.55	0.54
Gradient Boosting Classifier	max_depth = 25, n_estimators = 100	0.50	0.50
Ridge Classifier	alpha = 0.5	0.50	0.48
K Nearest Neighbours	weights = distance	0.46	0.46
Passive Aggressive Classifier	C = 0.5	0.46	0.43
Gaussian Naïve Bayes	var_smoothing = 1e-09	0.45	0.43
Gaussian Process Classifier	optimiser = fmin_l_bfgs_b	0.44	0.44
Stochastic Gradient Descent	penalty = l2	0.42	0.42

Crystal Habit model framework study results.

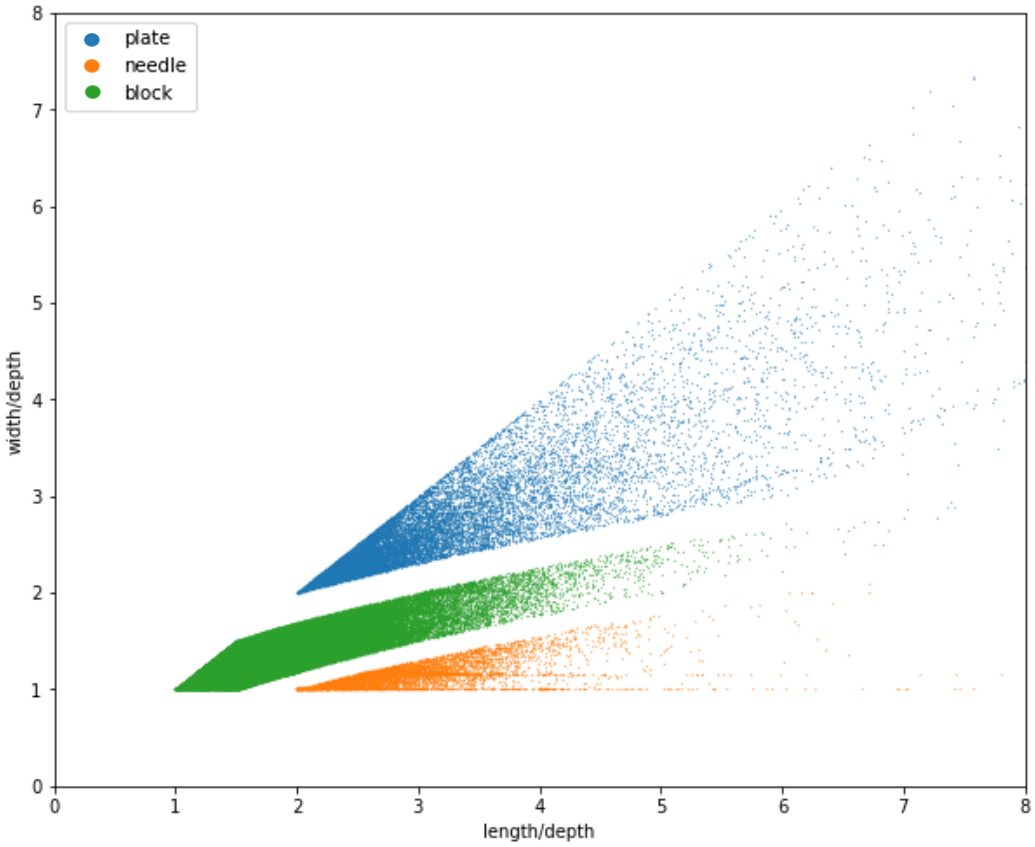
Estimator	Hyperparameters	Accuracy	F1 score
SVM Classifier	kernel = rbf, C = 10, gamma = 0.001	0.49	0.49
Random Forest Classifier	max_depth = 25, n_estimators = 100	0.49	0.48
Ridge Classifier	alpha = 1	0.47	0.47
Gradient Boosting Classifier	max_depth = 25, n_estimators = 100	0.44	0.43
K Nearest Neighbours	weights = distance	0.43	0.43
Gaussian Process Classifier	optimiser = fmin_l_bfgs_b	0.42	0.42
Passive Aggressive Classifier	C = 0.5	0.40	0.38
Stochastic Gradient Descent	penalty = l2	0.39	0.39
Gaussian Naïve Bayes	var_smoothing = 1e-09	0.39	0.37

Polymorphism framework study results.

Estimator	Hyperparameters	Accuracy	F1 score
SVM Classifier	kernel = rbf, C = 1, gamma = 0.001	0.62	0.62
Ridge Classifier	alpha = 0	0.60	0.60
Random Forest Classifier	max_depth = 25, n_estimators = 50	0.60	0.60
Gradient Boosting Classifier	max_depth = 25, n_estimators = 100	0.58	0.58
Gaussian Naïve Bayes	var_smoothing = 1e-09	0.56	0.54
Stochastic Gradient Descent	penalty = l1	0.55	0.55
K Nearest Neighbours	weights = uniform	0.54	0.54
Gaussian Process Classifier	optimiser = fmin_l_bfgs_b	0.54	0.54
Passive Aggressive Classifier	C = 0.5	0.52	0.51

C. Crystal habit labelling algorithm visualisation

Visual representation of the crystal habit classification rules applied to its corresponding dataset.



D. Extreme outlier threshold selection

Most statistical estimators and optimisation procedures assume features to have comparable scales. This requirement can be easily achieved through a simple data transformation procedure. However, chemical datasets are easily contaminated with outliers that can greatly affect the quality of the scaling process.

In early stages of this project, features were scaled to zero mean and unit variance. We rapidly noticed that when extreme outliers (points for which at least one of its feature values is extreme) were present in the original dataset, the scaling method (sensitive to outliers) would shrink all feature values to zero in order to include outliers within the new bounded feature space. This event leads to changes in the structure of the dataset and may compromise the patterns to be found in data, during exploratory data analysis and model fitting. The solution was to scale features to zero median and unit interquartile range. All features would be projected onto comparable scales, and given its robustness to outliers, an extreme value in the original dataset would still be extreme in the scaled output.

The advantage of removing extreme value outliers after scaling is that all features are on comparable scales and a single value threshold is required for outlier filtering. With all this in mind, we investigated the impact of adopting different extreme value thresholds on model performance. The crystallizability dataset was cleaned from incomplete entries and treated for domain outliers. Each feature was scaled to zero median and unit interquartile range, and data was split into training and test set (70/30). After scaling, all feature values are spread around zero but not necessarily in a symmetrical fashion.

The goal of this experiment was to find a rough cut off that 1) works for all features, 2) does not lead to the loss or change of the distributional characteristics of any feature, and 3) only exclude the most extreme and potentially harmful outliers, *i.e.* unjustifiably long feature distribution tails. Therefore, the training set was treated for extreme outliers, using different threshold values. Accordingly, all examples with at least one scaled feature value above a given threshold or below its symmetric value are discarded. Each version of the training set was balanced with respect to the target variable and used to train a predictive model. Refer to the first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model*** for more details on the crystallizability dataset and modelling framework.

All models were evaluated with a scaled version of the test set that was never treated for extreme outliers. Results are shown in the **Figure 27** below. Note that the x-axis

does not directly correspond to the extreme value threshold, but to the exponents of a base 10.

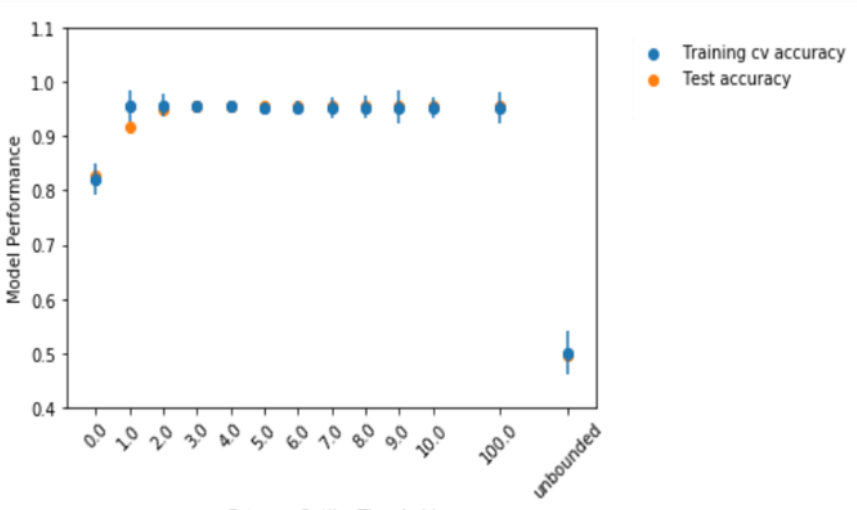


Figure 27 Extreme outlier threshold study results. Blue points represent the mean training accuracy value of a 3-fold cross validation routine, with error bars representing the standard deviation. Orange points represent test accuracy. The x-axis represents the exponent of a base 10, which together form the extreme outlier threshold. Furthermore, the extreme threshold mark in the x-axis of 100 is not represented at scale, whilst the 'unbounded' mark is figuratively represented in the plot and refers to a scenario where the training set is not treated for extreme outliers at all. These results refer to the crystallizability dataset and model. Visit the first paragraph of ***Chapter 4 / Methods / Crystallizability Dataset & Model*** for more details.

We speculate that the lower model performance obtained for very low extreme outlier thresholds (1 or 10^0) stems from the fact that not enough examples where available during the training procedure. Indeed, strict thresholds may discard most of the data and if not enough is available during training, the success of the mapping task can be compromised. On the other hand, harmful extreme values are allowed in the training set when very large or no thresholds are adopted. These values are so extreme that end up compromising the quality of the model learning process as well. An in between range of suitable threshold values is identified. All these lead to a maximal model performance on the test set. Amongst these, smaller thresholds are stricter and lead to smaller dataset sizes. Despite the decrease in the number of examples, we found that our models would not suffer any performance drop. Furthermore, smaller thresholds have also been found to guarantee the stability of the

multiple methods implemented throughout this project. Thus, we adopted an extreme value threshold of 100 (10^2).

Note that the extreme outlier discarding method does not claim to remove all univariate outliers from a dataset. Thus, thresholds should always be large enough to avoid losses or changes of the main characteristics of the feature value distribution. The method is meant to simple remove unjustifiably long tails, since only really extreme values seem to be harmful to the model training process. Training sets have been treated for extreme outliers since the adoption of the robust scaler.

E. Functional groups, α, and β values currently implemented in the IIP library

This is information retrieved from *Hunter et al.*[149]

Functional Group	α	β
alkane	0.4	0.3
ketone	1.5	5.8
carbamate	2.8	7.3
alkene	0.7	1.1
sulfonamide	2.8	5.8
ether	0.9	5.3
aldehyde	1.6	4.7
amide	2.9	8.3
furan	1.7	2.2
thioether	1	3.6
thiol	1.7	2.7
pyrrole	3	4.1
alkyl_iodide	0	2.2
sulfinamide	3.2	8.3
alkyl_fluoride	1.2	2.7
thiophenol	1.8	2.2
thioamide	3.3	5.8
alkyl_chloride	1.3	2.2
nitroalkane	0	3.7
carboxylic_acid	3.6	5.3
alkyl_bromide	1.3	2.2
aryl_chloride	1.3	1.6
aryl_fluoride	1.4	1.6

Functional Group	α	β
aryl_bromide	0	1.6
aryl_iodide	0	1.6
alkyne	1.9	2.7
ester	1.5	5.3
amine	1.5	7.8
imine	0	5.8
sulfone	0	6.3
disulfide	0	3.7
phosphinate_diester	0	8.9
sulfate_diester	0	4.7
nitrile	0	4.7
sulfoxide	0	8.9
isothiocyanate	0	2.7
thiocyanate	0	2.7
amidine	0	8.9
phosphine_oxide	0	9.9
benzene	1	2.2
urea	3	8.3
imidazole	3.7	0
phenol	3.8	2.7
aniline	2.1	5.3
pyridine	1.4	7
alcohol	2.7	5.8

F. Feature redundancy studies

Feature redundancy is measured using the adjusted coefficient of determination (R^2), which represents the proportion of variance of a continuous dependent variable that the independent counterparts can explain under a given statistical modelling framework.[161] Its adjustment accounts for its inflation when the number of explanatory variables increases.[160] Note that the R^2 coefficient is often restricted to the interval 0 and 1, but negative values can occur when a non-linear function is used to fit the data.[299]

HBND results.

IIP Feature	Mean R^2
a_VSA4	1.0 +/- 0.0
total_alpha	0.894 +/- 0.005
total_beta	0.882 +/- 0.003
beta_vsa_norm	0.75 +/- 0.007
alpha_vsa_norm	0.744 +/- 0.011
beta_mw_norm	0.742 +/- 0.007
alpha_mw_norm	0.739 +/- 0.01
average_beta	0.707 +/- 0.007
b_VSA2	0.702 +/- 0.009
a_VSA2	0.68 +/- 0.031
b_VSA3	0.619 +/- 0.008
average_alpha	0.591 +/- 0.022
b_VSA1	0.574 +/- 0.008
a_VSA0	0.563 +/- 0.006
a_VSA1	0.55 +/- 0.007
b_VSA4	0.524 +/- 0.015
b_VSA0	0.519 +/- 0.018
max_beta	0.516 +/- 0.007
a_VSA3	0.482 +/- 0.015
max_alpha	0.456 +/- 0.02
min_beta	0.445 +/- 0.016
min_alpha	0.393 +/- 0.015
beta_logp_norm	-0.019 +/- 0.002
alpha_logp_norm	-0.02 +/- 0.003

Crystal Habit results.

IIP Feature	Mean R^2
a_VSA4	1.0 +/- 0.0
total_alpha	0.866 +/- 0.025
alpha_mw_norm	0.851 +/- 0.007
alpha_vsa_norm	0.849 +/- 0.006
total_beta	0.846 +/- 0.025
beta_mw_norm	0.831 +/- 0.004
beta_vsa_norm	0.828 +/- 0.003
a_VSA2	0.768 +/- 0.011
average_beta	0.759 +/- 0.002
average_alpha	0.746 +/- 0.013
b_VSA2	0.745 +/- 0.012
max_alpha	0.744 +/- 0.004
b_VSA1	0.713 +/- 0.022
b_VSA0	0.701 +/- 0.003
a_VSA0	0.682 +/- 0.018
max_beta	0.679 +/- 0.004
b_VSA3	0.678 +/- 0.013
a_VSA1	0.676 +/- 0.006
min_alpha	0.534 +/- 0.007
min_beta	0.443 +/- 0.011
alpha_logp_norm	0.121 +/- 0.005
beta_logp_norm	0.117 +/- 0.002
b_VSA4	-5.267 +/- 2.781
a_VSA3	-11.34 +/- 2.486

Polymorphism results.

IIP Feature	Mean R^2
a_VSA4	1.0 +/- 0.0
total_alpha	0.731 +/- 0.023
total_beta	0.719 +/- 0.011
alpha_mw_norm	0.704 +/- 0.005
alpha_vsa_norm	0.702 +/- 0.006
max_alpha	0.673 +/- 0.013
beta_mw_norm	0.662 +/- 0.014
beta_vsa_norm	0.653 +/- 0.016
average_beta	0.619 +/- 0.021
average_alpha	0.589 +/- 0.022
max_beta	0.582 +/- 0.023
b_VSA2	0.542 +/- 0.02
a_VSA1	0.424 +/- 0.027
b_VSA1	0.414 +/- 0.017
a_VSA0	0.365 +/- 0.061
b_VSA3	0.326 +/- 0.028
min_alpha	0.3 +/- 0.013
b_VSA0	0.27 +/- 0.068
min_beta	0.23 +/- 0.02
a_VSA2	0.007 +/- 0.18
alpha_logp_norm	-0.021 +/- 0.013
beta_logp_norm	-0.029 +/- 0.005
a_VSA3	-0.813 +/- 0.217
b_VSA4	-3.299 +/- 0.353

G. Feature relevance studies

Only IIP features with feature redundancy results satisfying $R^2 < 0.5$ were considered for feature relevance studies. Feature relevance values correspond to the difference in model performance (mean F1 score of five 2-fold cross validation routines, based on the original statistical frameworks initially used for property prediction) when each unique feature is or is not added to the original dataset for material property modelling. The statistical significance of the difference in model performance was evaluated with a Wilcoxon sign-ranked test and a two-tail p-value cut-off of 0.05.

HBND results.

IIP Feature	Difference in Performance	p value	Significant?
max_alpha	0	0.878	no
min_beta	-0.001	0.959	no
min_alpha	-0.001	0.878	no
beta_logp_norm	0.002	0.241	no
alpha_logp_norm	0.002	0.386	no
a_VSA3	0	0.612	no

Crystal Habit results.

IIP Feature	Difference in Performance	p value	Significant?
min_beta	-0.002	0.139	no
alpha_logp_norm	0.004	0.013	**yes**
beta_logp_norm	0.005	0.017	**yes**
b_VSA4	0	0.579	no
a_VSA3	0	0.579	no

Polymorphism results.

IIP Feature	Difference in Performance	p value	Significant?
b_VSA1	0.001	0.634	no
a_VSA1	0.002	0.153	no
b_VSA3	0.002	0.262	no
min_alpha	-0.003	0.262	no
a_VSA2	0.002	0.297	no
b_VSA0	-0.001	0.348	no
a_VSA3	0.004	0.083	no
min_beta	-0.001	0.878	no
alpha_logp_norm	0.002	0.262	no
beta_logp_norm	0	0.508	no
b_VSA4	0	1	no
a_VSA0	0	1	no

H. List of *in-built* multivariate outlier detection estimators

1. Angle-Based Outlier Detection (probabilistic). It measures the variance of a sample's weighted cosine scores to all its neighbours.[242]

2. Isolation Forest (outlier ensemble). It measures the average depth of a tree-based structure that it takes to isolate a given sample, over a forest of random trees. [238,239]

3. Local Outlier Factor (proximity-based). It measures the local deviation of the density of a given sample with respect to its neighbours.[240]

4. One class SVM (linear model). It detects the soft boundary of the normal set of examples and measures the signed distance to the separating hyperplane.[241]

5. Histogram-Based Outlier Detection (proximity-based). It calculates the density of the bin to which each feature value of the multivariate point is assigned to and measures the sum of the log of the densities for all features.[243]

6. Mahalanobis- Based Outlier Detection (proximity-based). It measures the Mahalanobis distance of a sample to a defined centroid.[244]

Multivariate outlier detection methods are fitted to the training set with a contamination factor of 0.01 to exclude any potential multivariate outliers present in the training data. All other hyperparameters are set to the pyod library default values. For consistency, outliers are assigned with larger anomaly scores.

I. Generalisability studies

Crystallizability

MOD Model	10th qt	25th qt	50th qt	75th qt	90th qt	Unbounded
abod	**0.98**	**0.96**	**0.96**	**0.95**	**0.95**	**0.95**
mahalanobis	0.96	0.95	0.95	0.95	0.94	0.95
hbos	0.96	0.95	0.95	0.94	0.94	0.95
if	0.97	0.95	0.95	0.94	0.95	0.95
lof	0.97	0.96	0.96	0.95	0.95	0.95
ocsvm	0.94	0.95	0.95	0.94	0.94	0.94

Internal model performance (F1 score): 0.95 +/- 0.02
Selected multivariate outlier detection model: abod
Selected generalisability (anomaly) threshold: 90th quantile of the training anomaly score distribution

HBND

MOD Model	10th qt	25th qt	50th qt	75th qt	90th qt	Unbounded
abod	**0.63**	**0.62**	**0.57**	**0.55**	**0.55**	**0.55**
mahalanobis	0.49	0.47	0.52	0.53	0.54	0.55
hbos	0.53	0.57	0.53	0.53	0.52	0.52
if	0.55	0.54	0.54	0.55	0.55	0.55
lof	0.52	0.53	0.50	0.54	0.55	0.55
ocsvm	0.53	0.51	0.53	0.54	0.54	0.54

Internal model performance (F1 score): 0.55 +/- 0.01%
Selected multivariate outlier detection model: abod
Selected generalisability (anomaly) threshold: 50th quantile of the training anomaly score distribution

Crystal Habit

MOD Model	10th qt	25th qt	50th qt	75th qt	90th qt	Unbounded
abod	0.64	0.57	0.50	0.49	0.50	0.50
mahalanobis	0.47	0.51	0.50	0.49	0.51	0.51
hbos	0.61	0.56	0.53	0.49	0.51	0.51
if	0.50	0.49	0.49	0.49	0.48	0.48
lof	0.66	0.53	0.51	0.50	0.50	0.50
ocsvm	0.41	0.49	**0.52**	**0.51**	**0.51**	**0.49**

Internal model performance (F1 score): 0.49 +/- 0.02 %
Selected multivariate outlier detection model: ocsvm
Selected generalisability (anomaly) threshold: 50th quantile of the training anomaly score distribution

Polymorphism

MOD Model	10th qt	25th qt	50th qt	75th qt	90th qt	Unbounded
abod	0.67	0.59	0.58	0.59	0.60	0.61
mahalanobis	0.44	0.42	0.54	0.57	0.59	0.59
hbos	0.44	0.46	0.52	0.58	0.59	0.60
if	0.50	0.59	**0.63**	**0.63**	**0.61**	**0.60**
lof	0.54	0.58	0.57	0.60	0.60	0.61
ocsvm	0.43	0.49	0.60	0.57	0.60	0.60

Internal model performance (F1 score): 0.62 +/- 0.02
Selected multivariate outlier detection model: if
Selected generalisability (anomaly) threshold: 75th quantile of the training anomaly score distribution

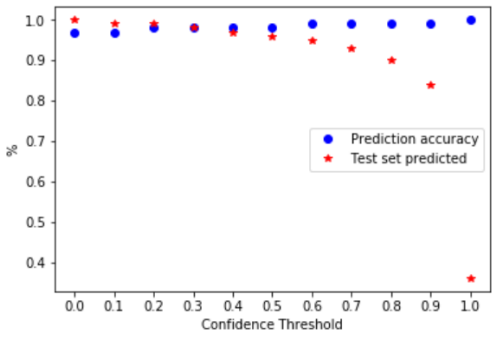
J. Confidence studies and tool applications

Crystalisabillity

1. Robustness studies (ct=0.05)
Number of molecules being assigned to a given class with only <5% difference between the top two probability estimates: 476

2. Study of the effect of increasing confidence thresholds on the fraction of test examples that a model can predict with confidence and the corresponding accuracy.

Maximum ct enabling meaningful statistics: 1.00



Conf. Threshold	% Test set Evaluated	Accuracy
0.00	1.00	97%
0.05	1.00	97%
0.50	0.96	97%
1.00	0.36	100%

3. Predict the test set over seven classification rounds of decreasing confidence thresholds.

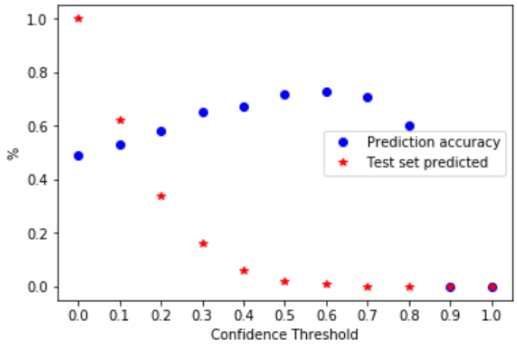
Round	Confidence Threshold	Confident Predictions	Correct Predictions	Accuracy (%)
1	1.0	47959	47911	100
2	0.9	66705	66064	99
3	0.8	8137	7722	95
4	0.7	3568	3204	90
5	0.6	2253	1919	85
6	0.5	1713	1379	81
7	0.4	1335	1023	77
8	0.3	1125	756	68
9	0.2	1018	644	63
10	0.1	1027	598	58
11	0.0	924	503	54

Crystal habit

1. Robustness studies (ct=0.05)
Number of molecules being assigned to a given class with only <5% difference between the top two probability estimates: 1,271

2. Study of the effect of increasing confidence thresholds on the fraction of test examples that a model can predict with confidence and the corresponding accuracy.

Maximum ct enabling meaningful statistics: 0.50



Conf. Threshold	% Test set Evaluated	Accuracy
0.00	1.00	49%
0.05	0.79	51%
0.30	0.16	65%
0.50	0.02	72%

3. Predict the test set over seven classification rounds of decreasing confidence thresholds.

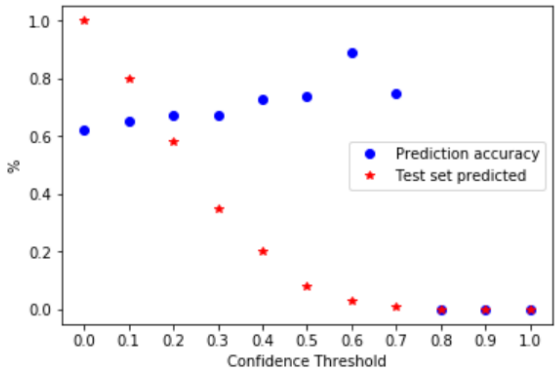
Round	Confidence Threshold	Confident Predictions	Correct Predictions	Accuracy (%)
1	0.5	110	87	79
2	0.4	211	153	73
3	0.3	505	305	60
4	0.2	1025	554	54
5	0.1	1521	738	49
6	0.0	2286	934	41

Polymorphism

1. Robustness studies (ct=0.05)
Number of molecules being assigned to a given class with only <5% difference between the top two probability estimates: 62

2. Study of the effect of increasing confidence thresholds on the fraction of test examples that a model can predict with confidence and the corresponding accuracy.

Maximum ct enabling meaningful statistics: 0.40



Conf. Threshold	% Test set Evaluated	Accuracy
0.00	1.00	62%
0.05	0.91	64%
0.20	0.58	67%
0.40	0.20	73%

3. Predict the test set over seven classification rounds of decreasing confidence thresholds.

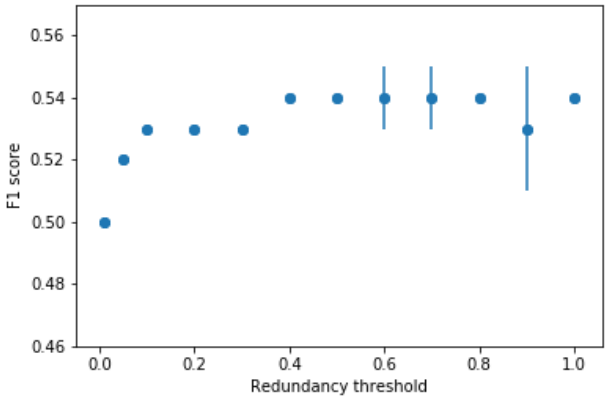
Round	Confidence Threshold	Confident Predictions	Correct Predictions	Accuracy (%)
1	0.4	143	104	73
2	0.3	106	64	60
3	0.2	163	109	67
4	0.1	162	94	58
5	0.0	142	76	54

K. Redundancy studies

Crystallizability

Redundancy Level	Dimensionality	Internal model performance
0.01	9	0.85 +/- 0.01
0.05	26	0.90 +/- 0.00
0.10	45	0.93 +/- 0.00
0.20	64	0.94 +/- 0.00
0.30	74	0.94 +/- 0.00
0.40	84	0.95 +/- 0.00
0.50	91	0.95 +/- 0.00
0.60	98	0.95 +/- 0.00
0.70	104	0.95 +/- 0.01
0.80	107	0.95 +/- 0.00
0.90	109	0.95 +/- 0.00
0.99	110	0.95 +/- 0.01

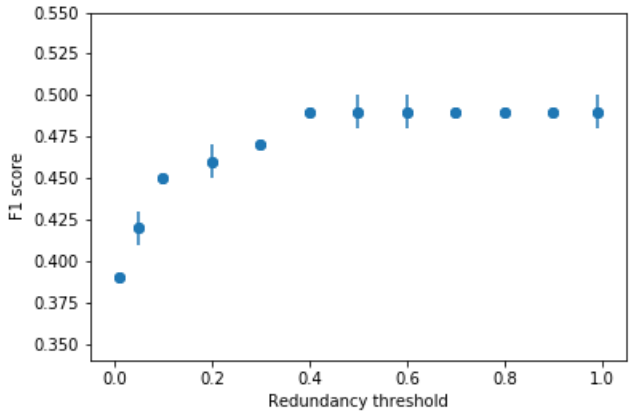
HBND



Redundancy Lvl	Dimensionality	Internal model performance
0.01	7	0.50 +/- 0.00
0.05	25	0.52 +/- 0.00
0.10	46	0.53 +/- 0.00
0.20	59	0.53 +/- 0.00
0.30	73	0.53 +/- 0.00
0.40	84	0.54 +/- 0.00
0.50	90	0.54 +/- 0.00
0.60	96	0.54 +/- 0.01
0.70	102	0.54 +/- 0.01
0.80	107	0.54 +/- 0.00
0.90	109	0.53 +/- 0.02
0.99	110	0.54 +/- 0.00

Redundancy level selected: 0.1

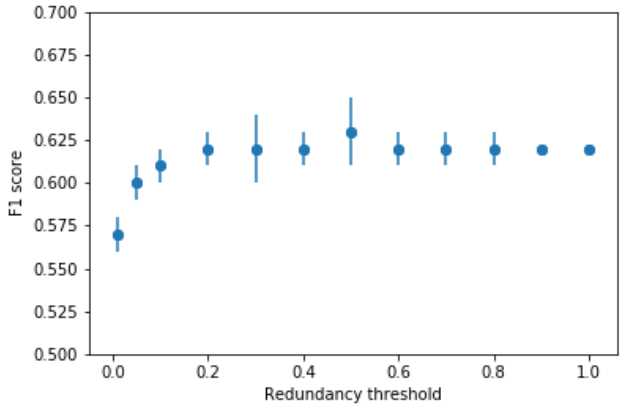
Crystal habit



Redundancy Lvl	Dimensionality	Internal model performance
0.01	6	0.39 +/- 0.00
0.05	21	0.42 +/- 0.01
0.10	40	0.45 +/- 0.00
0.20	60	0.46 +/- 0.01
0.30	75	0.47 +/- 0.00
0.40	85	0.49 +/- 0.00
0.50	91	0.49 +/- 0.01
0.60	99	0.49 +/- 0.01
0.70	103	0.49 +/- 0.00
0.80	108	0.49 +/- 0.00
0.90	109	0.49 +/- 0.00
0.99	110	0.49 +/- 0.01

Redundancy level selected: 0.4

Polymorphism



Redundancy Lvl	Dimensionality	Internal model performance
0.01	6	0.57 +/- 0.01
0.05	19	0.60 +/- 0.01
0.10	35	0.61 +/- 0.01
0.20	58	0.62 +/- 0.01
0.30	73	0.62 +/- 0.02
0.40	79	0.62 +/- 0.01
0.50	89	0.63 +/- 0.02
0.60	98	0.62 +/- 0.01
0.70	103	0.62 +/- 0.01
0.80	107	0.62 +/- 0.01
0.90	109	0.62 +/- 0.00
0.99	110	0.62 +/- 0.00

Redundancy level selected: 0.2

L. Evolution of the dataset's dimensionality and predictive power, as it flows through the feature selection pipeline

HBND

Feature set	*Dimensionality*	*Internal model performance*
Original	113	0.55 +/- 0.01
Non-redundant	46	0.53 +/- 0.00
Critical	10	0.50 +/- 0.00

Crystal habit

Feature set	*Dimensionality*	*Internal model performance*
Original	113	0.49 +/- 0.02
Non-redundant	85	0.49 +/- 0.00
Critical	10	0.44 +/- 0.00

Polymorphism

Feature set	*Dimensionality*	*Internal model performance*
Original	113	0.62 +/- 0.02
Non-redundant	58	0.62 +/- 0.01
Critical	10	0.59 +/- 0.00

M. Top most relevant features driving the prediction problem

HBND

Feature name	*Description*
NHOHCount	Number of NH and OH groups
NOCount	Number of N and O atoms
FractionCSP3	Fraction of sp3 hybridised carbons. Mainly covers flexibility.
PEOE_VSA10	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions
VSA_EState7	MOE-type descriptors using surface area and Estate indices contributions
PEOE_VSA4	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions
Kappa1	Molecular shape index for one bonded fragment
EState_VSA6	MOE-type descriptors using Estate indices and surface area contributions.
EState_VSA2	MOE-type descriptors using Estate indices and surface area contributions.
VSA_Estate8	MOE-type descriptors using surface area and Estate indices contributions

Crystal habit

Feature name	*Description*
PEOE_VSA10	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions
SlogP_VSA6	MOE-type descriptors using Wildman-Crippen log P and surface area contributions.
SMR_VSA10	MOE-type descriptors using Wildman-Crippen molecular refractivity and surface area contributions.
BalabanJ	Topological index J based on the distance sums as a graph of a molecule
SlogP_VSA3	MOE-type descriptors using Wildman-Crippen log P and surface area contributions.
Chi4n	Molecular valence connectivity index for a path of order 4
Kappa3	Molecular shape index for three bonded fragments
TPSA	Topological polar surface area
PEOE_VSA4	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions
PEOE_VSA9	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions

Polymorphism

Feature name	*Description*
Ipc	Information content of the coefficients of the characteristic polynomial of the adjacency matrix of a hydrogen-suppressed graph of a molecule
NumHDonors	Number of hydrogen bond donor groups
NumHAcceptors	Number of hydrogen bond acceptor groups
PEOE_VSA11	MOE-type descriptor accounting for partial charge and van der Waals surface area contributions
MolWt	Molecular Weight. Mainly covers size.
SMR_VSA4	MOE-type descriptors using Wildman-Crippen molecular refractivity and surface area contributions.
EState_VSA2	MOE-type descriptors using Estate indices and surface area contributions.
EState_VSA7	MOE-type descriptors using Estate indices and surface area contributions.
VSA_Estate7	MOE-type descriptors using surface area and Estate indices contributions
SlogP_VSA8	MOE-type descriptors using Wildman-Crippen log P and surface area contributions.