Improving QSAR model performance by relaxing linearity using smoothing splines within GAMs: Illustrated using anti-Streptococcus activity of prenylated phenolics

Mastwijk, Roos

Universiteit
Leiden
The Netherlands

Improving QSAR model performance by relaxing linearity using smoothing splines within GAMs

Illustrated using anti-*Streptococcus* activity of prenylated phenolics

Roos Mastwijk

Thesis advisor: Dr. J.A. Hageman

Defended on 20th March, 2024

MASTER THESIS

STATISTICS AND DATA SCIENCE
UNIVERSITEIT LEIDEN

Abstract

This study focussed on the development of a QSAR model for describing and predicting the anti-*Streptococcus* activity of prenylated phenolics. Traditionally, QSAR modelling is done with Multiple Linear Regression (MLR). Hereby, the relation between activity and molecular descriptors is assumed to be linear, which is not always fair in QSAR. Smoothing splines is a nonparametric method that relaxes the linearity assumption. Their use within Generalized Additive Models (GAMs) allows for multiple predictors. MLR models and GAMs with smoothing splines were fitted to a dataset with 28 prenylated phenolics experimentally tested against *Streptococcus mutans*. Inclusion of imputed Minimum Inhibitory Concentration (MIC) values for inactive molecules, addition of molecules from the Chalcone subclass and complementation with literature data that allowed for different bacterial strains and species did not benefit the models. The best MLR model that functioned as baseline was statistically not compliant ($R^2_{adjusted}$ = 0.586, Q^2_{LOO} = 0.358). Identification of high leverage molecules and their removal from the Applicability Domain (AD) did not improve the model performance. A GAM with a good fit ($R^2_{adjusted}$= 0.638) and sufficient internal predictive power (Q^2_{LOO}= 0.525) was obtained. The removal of three influential molecules from the AD improved the GAM fit further ($R^2_{adjusted}$ = 0.733). Internally, this model predicted well the activity against *S. mutans* from the hydrophobic volume at an interaction energy of -0.2 kcal/mol (Q^2_{LOO} = 0.607).

Contents

1 Introduction

Antimicrobial resistance poses one of the biggest threats to global healthcare (World Health Organization, 2021). It occurs when microorganisms adapt such that the effectivity of antimicrobial substances designed to kill them decreases. This phenomenon is mainly driven by the misuse and overuse of traditional antibiotics. The antimicrobial resistance induces challenges to the treatment of infections, leading to severe illness or even decease (Murray et al., 2022). Hence, novel antibacterial compounds are being investigated on their ability to combat infections.

The antimicrobial resistance problem is clearly illustrated by *Streptococcus mutans* (*S. mutans*), an oral bacterium that forms the major cause of dental caries. Fluoride is traditionally used to inhibit the growth of this bacterium and therefore acts as an anti-caries agent. However, *S. mutans* develops resistance against the antimicrobial effects of fluoride (Liao et al., 2017). Alternative antimicrobials are thus needed to prevent tooth decay.

Several studies have shown a that prenylated phenolics have promising antibacterial activity against *Streptococcus* and other pathogens. Prenylated phenolics are bioactive molecules found in high concentrations in edible plants such as licorice root and soybean (Chang et al., 2021). These prenylated phenolics all have a similar main structure, but the substituents they carry can differ widely (Balasundram et al., 2006). The differences in molecular structure cause a different Minimum Inhibitory Concentration (MIC), which is the minimum concentration of a molecule needed to stop bacterial growth (Kowalska-Krochmal & Dudek-Wicher, 2021). To determine if a prenylated phenolic is an effective antimicrobial, the relation between its structural properties and antibacterial activity is of interest.

Traditionally, Quantitative Structure-Activity Relationship (QSAR) models describe the relation between molecular descriptors and the MIC (Roy et al., 2015). Well performing and robust QSAR models make an excellent screening tool to save time and resources in synthesis and testing of molecules. Not only are QSAR models capable of predicting the expected MIC for a molecule, but they also acquire insights into the structural features needed for a desired MIC. The latter allows for synthesis of molecules with optimized activity.

QSAR modelling encounters particular challenges. Data used for QSAR modelling generally consist of more molecular descriptors than samples. Due to cost and time constraints, there is often limited capacity to test molecules in the lab causing only a small number of samples to be available. The molecular descriptors are however easily calculated and many have been invented. This is problematic in most statistical models as it leads to overfitting and consequently poor predictive accuracy. Feature selection eliminates this problem by selecting the most relevant descriptors. Additionally, feature selection helps keeping the model parsimonious and interpretable (Hageman, 2022).

The most frequently used regression model in QSAR is Multiple Linear Regression (MLR) (Liu & Long, 2009). MLR assumes a linear relation between the molecular descriptors and the MIC resulting in a straightforward and highly interpretable model. However, one should be careful with assuming linearity and check if this assumption holds. If not, linear methods neglect information in the data leading to decreased accuracy. Relaxing linearity by methods such as smoothing splines and Generalized Additive Models (GAMs) can be considered. Smoothing splines are very flexible and can model many different nonlinearities without assuming a functional form. The use of smoothing splines within GAMs allows for multiple predictor models while interpretability is maintained as additivity remains. Since there are limited studies available on the use of nonlinear methods in QSAR, these properties make the combination of smoothing

splines and GAMs a suitable exploratory method.

A recurring problem in QSAR is that molecules can be very diverse in nature and exhibit different modes of action. Molecules with very different modes of action cannot be combined in a single model. A concept used to investigate and define which molecules can be considered for a model is the Applicability Domain (AD) (Weaver & Gleeson, 2008). The AD is defined by molecules that show consistent behaviour indicated by small residuals and the absence of high leverage molecules. Molecules showing high residuals or high leverages will be removed from the data before the model is fit. However, calculating leverages for nonlinear models is non-trivial.

The aim of this thesis project is twofold. The first aim is to obtain an accurate and interpretable model that can accurately predict the MIC of prenylated phenolics against *Streptococcus* in datasets provided by the Food Chemistry department of Wageningen University and Research. The second aim is to investigate if QSAR model performance can be improved by relaxing linearity using smoothing splines within GAMs. As a third aim, the effect of complementation with literature data and imputation of MIC values for inactive prenylated phenolics on the statistical models will be reviewed.

To address the research aims, forward-MLR models and forward-GAMs are fitted to an experimental dataset with and without imputed MIC values and two datasets augmented with literature data. The forward-MLR models function as a baseline method to compare the performance of the forward-GAMs with. The AD for the forward-MLR models is determined in terms of leverage in a leave-one-out cross-validation (LOOCV) context. For the forward-GAMs, the AD is determined by assessing the LOOCV model performance to examine the influence of the individual prenylated phenolics. Lastly, the forward-MLR models and forward-GAMs are interpreted, the forward-GAMs by means of profile plots.

2 Materials and Methods

2.1 The dataset

The QSAR dataset employed in this study was collected experimentally and from literature by the Food Chemistry department of Wageningen University and Research. They assessed the anti-*Streptococcus* activity of prenylated phenolics originating from plant sources. For each prenylated phenolic in the dataset, the following variables are included:

- The compound name, subclass, and molecular weight
- Experimental conditions used:
  - The initial bacterial inoculum size
  - The assay method for determining the MIC
  - The incubation time, temperature, and atmosphere
  - Growth medium
- The bacterium: *S. mutans*, or another species of *Streptococcus*
- The bacterial strain
- The data source:
  - Experimental (determined in the lab)
  - From literature
- The antibacterial activity (MIC) in µg/mL and the pMIC (M)
- 440 molecular descriptors, calculated by Molecular Operating Environment (MOE).

The Minimum Inhibitory Concentration (MIC) is the lowest concentration of a molecule that results in a bacterial cell count equal to or lower than that of the initial inoculum size. A lower MIC indicates higher activity of the molecule. The pMIC was calculated from the MIC using the logarithmic transformation:

$$pMIC = -log_{10}(\,(MIC/molecular\ weight)/1000)\,).$$

The pMIC provides a more convenient scale for statistical modelling. Resulting from this transformation, the interpretation of the pMIC is opposite from the MIC: a higher pMIC indicates higher activity of the molecule.

Molecular descriptors were calculated from optimized chemical molecular structures by MOE software. The descriptors include both 2D and 3D molecular descriptors. Examples are descriptors on structural, electronic, steric, geometric, and physicochemical properties.

This study aims at modelling the pMIC as a function of the MOE descriptors only. The remaining variables were only used to gain insights into the differences in experimental conditions compared to other studies to make appropriate choices when selecting literature data. If multiple articles on a specific prenylated phenolic were available, the study with experimental conditions most resembling those in the Food Chemistry laboratory was selected.

Due to the combination of experimental and literature data, the total dataset was divided into three different sets. The first set contained the MICs for 42 prenylated phenolics (shown in figure 1) against the same *S. mutans* strain experimentally determined in the Food Chemistry laboratory. Activity of these prenylated phenolics was assessed

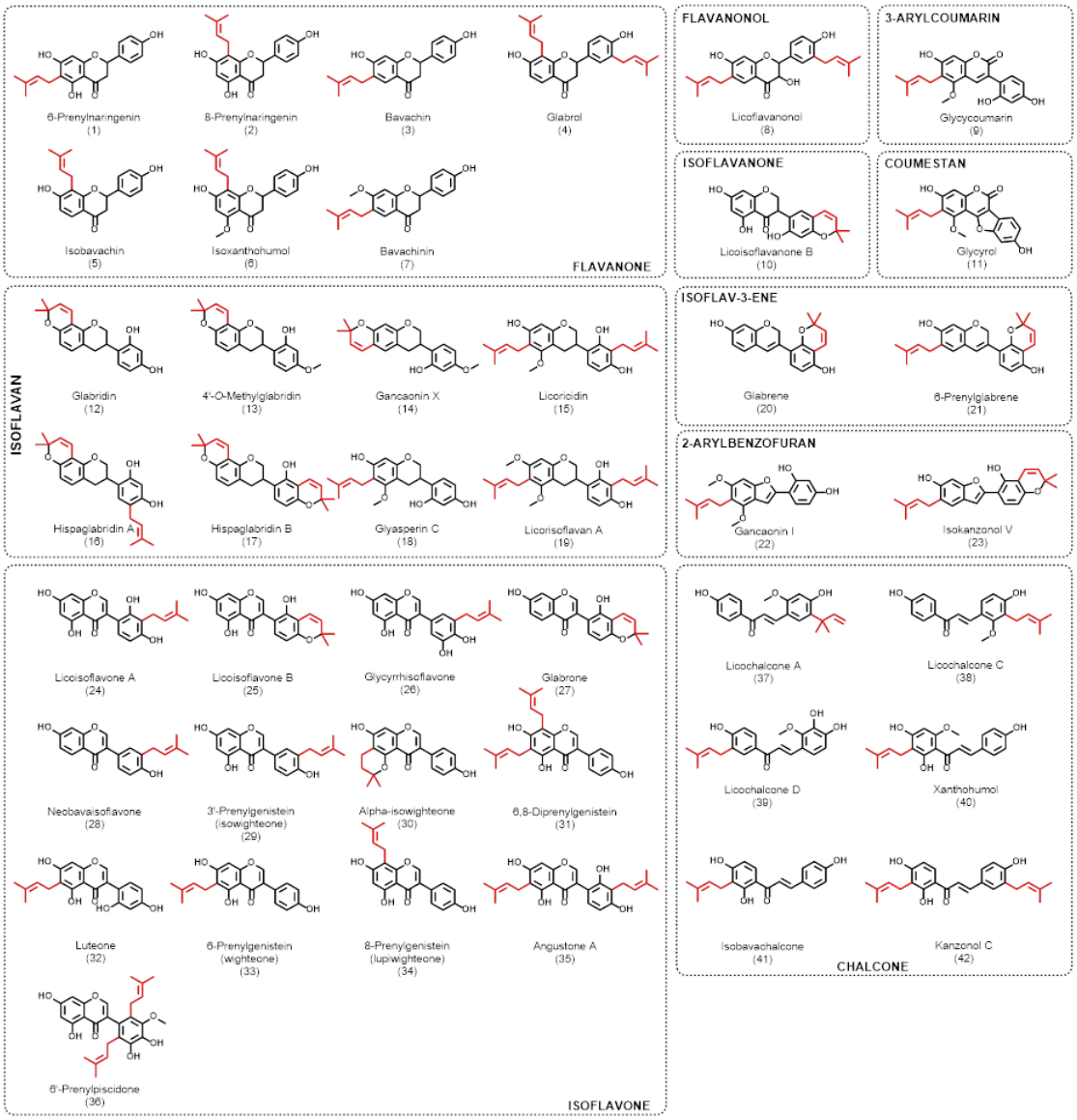

Figure 1: The 42 molecules that were experimentally tested in the Food Chemistry laboratory from Wageningen University and Research.

by broth microdilution assay. Here, equal volumes (100 µL) of bacterial inoculum (size 4.9 logCFU/mL) were added to a series of 2-fold dilutions (1.56 – 100 µg/mL) from each prenylated phenolic in growth medium TSB. The samples were incubated in an aerobic environment for 24 hours at 37°C. Every 10 minutes the bacterial growth was measured using Optical Density (OD) at 600nm. Bacterial growth causes a more turbid sample, thereby more scattering of light and thus a higher OD. In OD, the time to detection (TTD) refers to the time it takes to reach a detectable level of bacterial growth or inhibition. Here, the TDD threshold was set at 0.05 units change, meaning that an increase of 0.05 in 10 minutes indicated significant bacterial growth. Viable cell count was performed when the change in OD was below 0.05. The lowest concentration of the prenylated phenolic at which the cell count was equal to or lower than the initial inoculum size was determined as the MIC.

The second dataset comprised the experimental data augmented with literature data. Here, it was mandated that the bacteria in the literature were of the species *S. mutans*, while variation in the bacterial strains within that species was permitted.

39 additional prenylated phenolics were added to this dataset resulting in a total of 81 observations. The third set was further augmented with literature data that allowed for diverse *Streptococcus* species and varying strains. This resulted in a dataset with 122 prenylated phenolics.

Molecular descriptors with zero variance were removed from the data. A descriptor with the same value for each molecule cannot explain any difference between them. Inclusion of such predictors leads to problems in statistical modelling. Additionally, highly collinear variables were removed. Strongly correlated variables are problematic because they destabilize statistical models and reduce their predictive power.

2.2 Imputed MIC values for inactive prenylated phenolics

9 out of the 42 experimentally tested prenylated phenolics failed to inactivate *S. mutans* at their highest concentration (100 µg/mL). Using their TDD and the TTD of the negative control where *S. mutans* was inactivated by 2% DMSO, the MIC for these prenylated phenolics was estimated as:

1. TDD between 0 – 5 hours compared to the negative control:

   Estimated MIC = highest concentration x 10

2. TDD between 5 – 10 hours compared to the negative control:

   Estimated MIC = highest concentration x 5

3. TDD > 10 hours compared to the negative control:

   Estimated MIC = highest concentration x 2

2.3 R packages and functions

The statistical analysis for this study was performed in R. Table 1 provides and overview of the packages and functions used.

Table 1: Overview of R packages and functions used in this study.

Package	**Function**	**Used for:**
caret	findCorrelation	Identification of highly correlated predictors
dplyr	rename	Renaming column headers
ggplot2	ggplot	Visualization of pMIC observed versus pMIC predicted plot
ggrepel	geom_label_repel	Adding non-overlapping labels to ggplot figures
mgcv	gam	Fitting Generalized Additive Models
mgcv	s	Modelling predictors as smooth terms in Generalized Additive Models. The default type of smoothing splines used in this function is thin plate (Wood, 2003). The optimization of λ performed using the default Generalized Cross-Validation.

2.4 Multiple Linear Regression

MLR is a statistical modelling technique used to analyze and predict a quantitative outcome as a function of multiple predictor variables. The equation for MLR has the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon.$$

In this equation, Y is a vector containing the outcome values for each of the observations from the dataset. Vector X_j includes the values for the jth predictor and β_j are unknown coefficients describing the weight of each predictor on the outcome. The error term ϵ represents the unknown error for each observation. Thereby, the discrepancy between an observed outcome and the outcome that is predicted from an MLR model is quantified. The aim of this QSAR study is to predict the pMIC from molecular descriptors. In that context, coefficient β_j is interpreted as the change in pMIC upon a one unit increase in molecular descriptor X_j, while keeping all remaining molecular descriptors at a constant value.

Fitting an MLR model involves the estimation of the unknown coefficients β_1, β_2, $\ldots$, β_p. Given the regression coefficient estimates, the formula that can be used to make predictions on the outcome is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_p x_p.$$

The values for β_1, β_2, $\ldots$, β_p are estimated by minimizing the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - ... - \hat{\beta}_p x_{ip})^2.$$

This method of determining the β_j coefficients is called the least squares approach. Matrix algebra is used to calculate the estimates (Fox, 2015).

MLR is a parametric regression method since the functional form of the relation between the outcome and the predictors is predetermined. The relation is assumed to be linear. Due to this linearity assumption, the interpretation of the MLR equation is straightforward: a change in the outcome is a linear combination of the changes in predictors weighted by their respective coefficients. This also shows the additive nature of MLR as the linear combination is the sum of individual contributions of the weighted predictor values. The simplicity and interpretability resulting from the linearity assumption make MLR a very popular method in QSAR studies (Liu & Long, 2009).

Despite its benefits, the linearity assumption can be problematic in QSAR studies. Some relations between molecular descriptors and the pMIC are expected to show nonlinear behavior as linearity is rarely true, especially in biological context. For example, molecular descriptors can follow a parabola and have an optimum at a certain value, or a step function with a threshold for certain activity. Falsely assuming linearity causes MLR to ignore such information present in the data. In this light, it is expected that the relaxation of linearity improves QSAR model performance. Thus, it is important to check the linearity assumption (e.g. by examining Y versus X scatterplots) before applying MLR to a QSAR dataset and to adapt the modelling method if nonlinearities are observed.

2.5 Smoothing Splines

Discovery of nonlinear patterns in QSAR data makes it unfair to remain with the linearity assumption. Nonlinearities require a method that can capture more nuanced relationships by relaxing linearity. Selection of a method that correctly fits the nonlinear

relationship is important. Visual inspection of Y versus X scatterplots can provide insights into the type of relation present, but is often affected by e.g. complex forms, clustering, and outliers. Since nonlinear methods are not yet frequently applied in QSAR, nonparametric regression is preferred. Nonparametric holds that no assumptions are made on the functional form of the relation between the outcome and the predictors. This prevents ignorance of information and allows unexpected patterns in the data to be captured. One nonparametric method that can be used on quantitative data is smoothing splines. The explanation of smoothing splines in this section is based on the book 'An Introduction to Statistical Learning' by James et al. (2021).

Due to their flexibility, smoothing splines can capture a wide range of nonlinear patterns and reveal underlying structures of the data. This makes it a great method to explore the nonlinearities between the pMIC and molecular descriptors. To understand how smoothing splines can relax the linearity assumption, a step-by-step pathway towards smoothing splines is illustrated in figure 2.

Figure 2a shows simple linear regression as the starting point. This linear regression describes the relation between a quantitative outcome Y and a single quantitative predictor X as:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 describes the intercept and β_1 the slope of the line. The error term is indicated with the ϵ.

The linear regression model can be extended in two different directions: polynomial regression and a step function. To arrive at polynomial regression, higher order terms of the predictor are included. The higher order terms are obtained by raising predictor X to powers up to a certain degree. Resultingly, a global nonlinear fit to the data is attained. Figure 2b shows an example of a second-degree polynomial. A degree-d polynomial is mathematically described by:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_d X^d + \epsilon,$$

where β_0 describes the intercept. The remaining β-coefficients assign a weight to the first and higher order predictor terms.

Extension of linear regression towards a step function requires the division of predictor values into bins. The resulting qualitative predictor allows a constant value for the outcome Y in each of the bins. Thereby, a stepwise function introduces specificity within a certain range. The mathematical representation of a step function with bins created by cutting the range of X in points (called knots) $c_1, \ldots, c_k$ is of the form:

$$\begin{aligned} C_0(X) &= f(X < c_1) \\ C_1(X) &= f(c_1 \leq X < c_2) \\ &... \\ C_{K-1}(X) &= f(c_{K-1} \leq X < c_K) \\ C_K(X) &= f(c_K \leq X). \end{aligned}$$

Cutting the range of predictor X into k knots results in $k+1$ bins. An example of a step function created with two knots and thus three bins is illustrated in figure 2c. Usually, the constant defined within each bin is the mean of values within that bin. In that case, the function f around the values in a bin is defined as the sum of all predictor values corresponding to the observations in the bin divided by the number of observations in that bin.

Polynomial regression and the step function can be combined to fit separate, usually low-degree polynomials to each bin. This combination is named a piecewise polynomial.

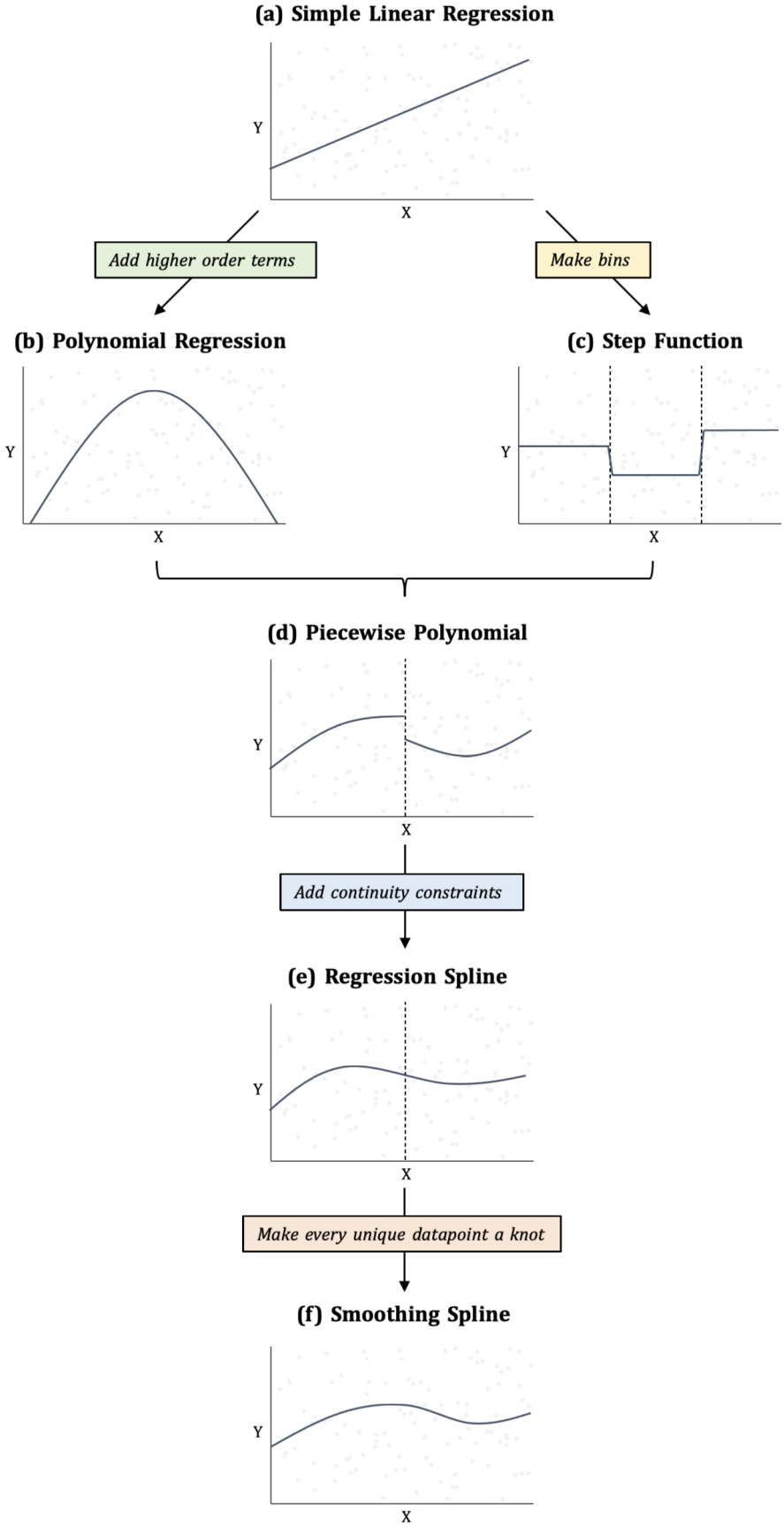

Figure 2: Step-by-step path diagram from simple linear regression towards a smoothing spline.

A piecewise third-degree polynomial, also called piecewise cubic polynomial, with a single knot in point c_1 is represented as:

$$Y = \begin{cases} \beta_{01} + \beta_{11}X + \beta_{21}X^2 + \beta_{31}X^3 + \epsilon & if\ X < c_1 \\ \beta_{02} + \beta_{12}X + \beta_{22}X^2 + \beta_{32}X^3 + \epsilon & if\ X \geq c_1. \end{cases}$$

The example of a piecewise polynomial with one knot illustrated in figure 2d shows that the individual polynomials do not connect at the knot. Despite the provided flexibility, this discontinuity is problematic for the interpretation and fit of the model. The remedy lies in the addition of constraints that provide connection and smoothness at the knot. Connection is obtained by imposing the constraint that the neighbouring endpoints of the polynomials should be of equal value at the knots. Smoothness at the knots is the result of continuity in the derivatives at each knot. A piecewise polynomial composed of degree-d polynomials with the constraint that its derivatives are continuous up to degree $d-1$ is called a degree-d regression spline. The derivatives are only continuous until degree $d-1$ as otherwise the resulting function becomes a global polynomial again. An example of a regression spline with one knot is shown in figure 2e.

The mathematical representations of the models up to piecewise polynomials were relatively straightforward. Due to the added constraints, the equation of a regression spline is rather complex. The constraints are included by means of truncated power basis functions. Truncated power basis functions only have value with a specific interval and are zero outside this interval. Mathematically, this is presented as:

$$g(x, \zeta) = (x - \zeta)_+^d = \begin{cases} (x - \zeta)^d & if\ x > \zeta \\ 0 & otherwise. \end{cases}$$

The restriction that the derivatives should be continuous up to degree $d-1$ imposes the addition of one truncated power function of degree d per knot. For a cubic regression spline with k knots, the resulting equation is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_{3+k}(X - \zeta_k)_+^3 + \epsilon.$$

The first four terms describe the global fit of the cubic regression spline. The truncated power basis functions provide a more nuanced fit to the global function on their interval. A cubic regression spline thus has $3 + k$ predictors, and a total of $4 + k$ regression coefficients are estimated. Resultingly, a cubic smoothing spline modelled using k knots uses $4 + k$ degrees of freedom.

How strong the truncated power basis functions influence the global function depends on their weights represented by the β-coefficients. This principle is illustrated in figure 3. The upper graph shows a global function in blue, and the lower graph a truncated power basis function in yellow. Adding this basis function to the global function shifts the curve of the global function at the specified interval which results in the green curve.

The progress from a regression spline to a smoothing spline lies in the number of knots. Choosing the right number of knots is a difficult task. Smoothing splines avoid this problem by not requiring a predefined number of knots to the create basis functions. Instead, each unique datapoint is considered a knot as illustrated in figure 2f. Consequently, the approach by which a smoothing spline is fitted differs from that of regression splines.

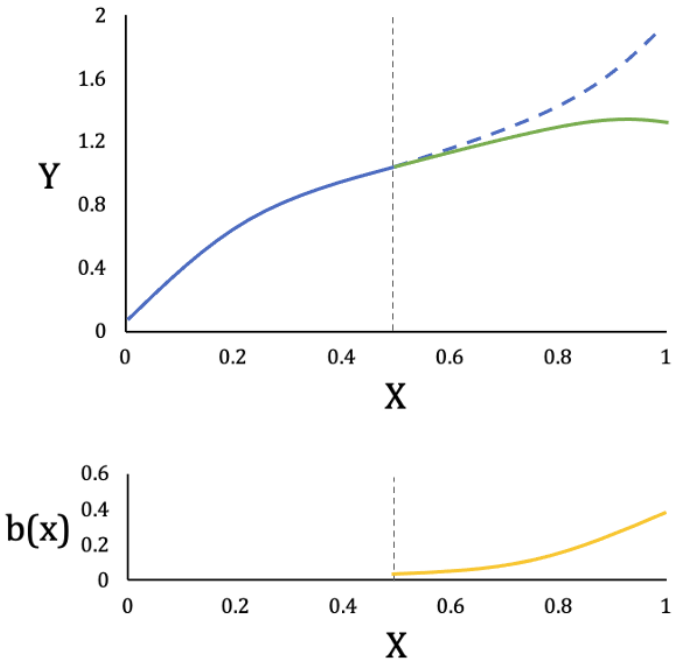

Figure 3: The addition of a truncated power basis function (lower graph, yellow) to a global function (upper graph, blue) that results in the piecewise polynomial (upper graph, green).

When fitting a curve to data, the goal is to find a function $g(x)$ that fits the data well. In other words, a function $g(x)$ should estimate the values of Y such that the difference with the observed Y is small. This corresponds to a low Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - g(x_i))^2.$$

Without any constraints on $g(x)$, this function can be chosen such that the RSS is zero, indicating a perfect fit to all datapoints. This phenomenon is called overfitting. Overfitting should be prevented as it leads to bad prediction models and decreased interpretability. So, function $g(x)$ should not only fit the data well but should also be a smooth function. Smoothing splines prevent overfitting by not minimizing the RSS, but minimizing a penalized RSS:

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt.$$

The added term is composed of two elements: tuning parameter λ and a measure of wigglyness described as the integral of the squared second derivative of function $g(x)$ at point t.

Understanding of the wigglyness measure requires knowledge on the concept of derivatives. The first derivative $g'(t)$ describes the slope of function $g(x)$ at point t. The second derivative $g''(t)$ describes the amount by which the slope changes at point t. When the second derivative at a point t is large it means that $g(x)$ drastically changes its slope in that point. Taking the integral of the squared second derivatives over the entire range at which $g(x)$ is defined can be said to provide a measure of wigglyness. Note that the value of $g''(t)$ is squared to prevent positive and negative changes in slope

from cancelling out. From this definition of wigglyness becomes intuitive that a very wiggly function has a frequently changing slope which corresponds to a large value for $\int g''(t)^2 dt$.

Part of the wigglyness of function $g(x)$ can be controlled by choosing the number of basis functions k instead of allowing k to equal the number of unique values of the predictor. When a predictor is modelled as smoothing spline, the degrees of freedom available for the basis functions equals k–1. Less degrees of freedom available decreases the number of basis functions that can be used and thus restricts the wigglyness.

Tuning parameter λ also controls the wigglyness as it encourages the function to be smooth by balancing the fit of the data and smoothness of the curve. Therefore, the aim is to optimize λ by using methods as Generalized Cross-Validation (GCV) or Restricted Maximum Likelihood (REML). When λ approaches zero, the measure of wigglyness becomes less influential. It allows the model to be flexible and wiggly, resulting in a curve closely fitting the datapoints. This comes with the risk of overfitting. Higher values for λ increase the importance of the wigglyness measure. It shifts the importance from fitting the data closely towards obtaining a smoother curve that is less prone to overfitting and able to reveal underlying trends. Figure 4 illustrates this effect of λ on two curves fitted on the same data. The final curvature resulting from the choice of k and the optimized λ is quantified in terms of the Effective Degrees of Freedom (EDF). Lower values for the EDF indicate smoother curves.

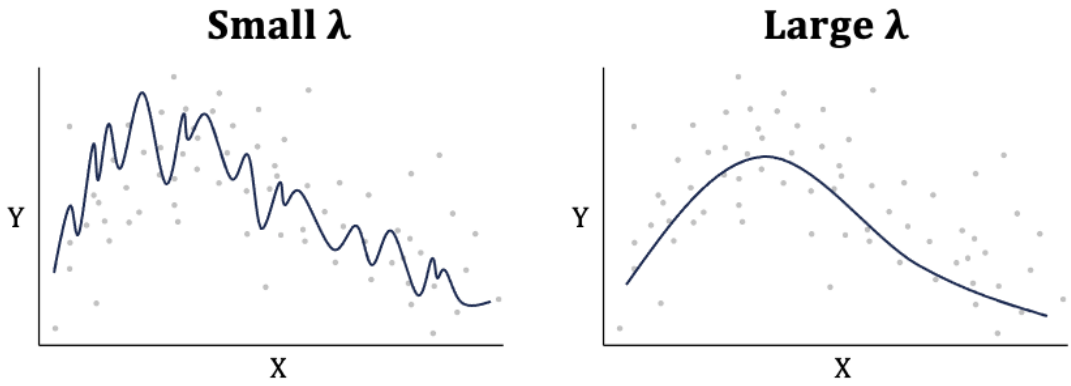

Figure 4: Smoothing spline fitted with a large λ (left) and a small λ (right).

Compared to other nonlinear modelling methods, smoothing splines remain relatively easy to interpret by means of profile plots. These plots provide a visual representation of the relationship between the outcome and a predictor while accounting for the smoothing effect. This intuitive nature of profile plots aids researchers to make informed interpretations about underlying pattern of the data. Figure 5 shows an example of a profile plot where the smooth function is plotted together with its confidence interval. The descriptor is represented on the x-axis and the y-axis represents the outcome values as determined by the smooth function. Profile plots always indicate the EDF corresponding to the curve. The small bars on the x-axis represent the observations at their descriptor value. Examination of the confidence interval gives an idea of the uncertainty associated with the smooth curve at different predictor values. From profile plots, the functional forms of relationships between the outcome and predictors can be observed.

A disadvantage of smoothing splines is that they can only model one single predictor at the time. In QSAR, individual molecular descriptors are not expected to be related strong enough to the pMIC such that they can sufficiently predict on their own. Generalized Additive Models (GAMs) can model multiple molecular descriptors as smoothing splines

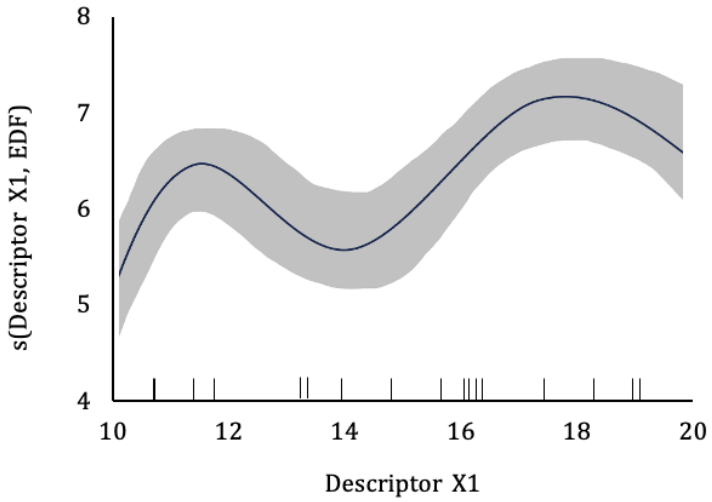

Figure 5: Example of a profile plot with its confidence interval. The descriptor values corresponding to the observations are indicated with bars on the x-axis. The y-axis represents the outcomes as modelled by the smooth function. The EDFs corresponding to the profile plot are indicated in the smooth term.

in one model.

2.6 Generalized Additive Models

Generalized Additive Models (GAMs) are used to fit an additive models with predictor terms that can be nonlinear. The general formula for a GAM is:

$$Y = f(X_1, X_2, ..., X_p) + \epsilon = \beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p) + \epsilon.$$

The functions f can be different for each predictor. For example, X_1 can be modelled as a linear term, X_2 as a third-degree polynomial and X_3 as a smoothing spline. In this study, all predictors are modelled as smoothing splines.

As described, modelling a single smoothing spline does not require the number of basis functions to be specified as all unique datapoints can be seen as knots. However, the number of basis functions can be set to a maximum that equals the number of unique values of the predictor. The same holds for the smooth terms in a GAM. This thesis aims to find the optimal GAM by varying the number of predictors included. It is possible that the first predictors included in the model claim all degrees of freedom. This hinders the inclusion of more predictors and comes with the risk on overfitting. Therefore, the number of basis functions for each predictor is rather specified. This study equally divides the total degrees of freedom amongst the predictors. The total degrees of freedom available equals the number of observations. An upper limit for the number of basis functions per predictor is determined as the total degrees of freedom divided by the number of predictors, always rounded downwards.

Each predictor included in the GAM can be interpreted using the same profile plots as illustrated in figure 5 of the previous paragraph on smoothing splines. However, the interpretation slightly changes due to the presence of multiple predictors in the model. The profile plots are now interpreted as the marginal effect of the considered predictor on the outcome while holding the other predictors at fixed values. Hereby, it is assumed that the predictors in the model are not correlated with each other to prevent misleading interpretations.

2.7 Forward selection

Extensive labor and high costs related to the purification and testing often results in only a limited number of molecules that are tested. On the other hand, molecular descriptors are easily calculated and many have been invented. This leads to QSAR data consisting of more molecular descriptors than molecules tested. The resulting high-dimensional data poses fundamental problems to statistical models. Firstly, models that use more predictors than observations do not have a unique solution. Next to that, not all predictors are expected to contribute to the outcome and the inclusion of many irrelevant predictors hinders the modelling. The flexibility and complexity of models with many predictors easily leads to overfitting, causing models to be fitted to the noise present in the data instead of to the underlying pattern. This reduces the predictive ability of the model. Also, a smaller set of predictors is desirable for model interpretation and benefits gaining insights into the underlying mechanisms (Hageman, 2022). To avoid these high-dimensionality problems, a variable selection method can be applied to build a model with only the most important predictors.

Forward selection is a straightforward way of selecting the most important predictors and is extensively used in QSAR. This method belongs to the wrapper methods as both the outcome and the predictors are considered in the selection process. The starting point of forward selection is a model with only the intercept. Predictors are added to the model one at the time, and for each predictor the improvement of the model is assessed. One measure that can be used to determine the model improvement is the Residual Sum of Squares (RSS). The RSS quantifies the unexplained variability in a model as the sum of squared differences between the actual values for the outcome and the values predicted by a statistical model. A smaller RSS implies that the predicted values are close to the actual values, indicating a better model fit. The predictor whose addition results in the lowest RSS is then selected into the model. The proceeding round then looks for the second-best predictor following the same procedure. How many rounds of forward selection are performed depends on the desired number of predictors to be selected or a criterium can be specified that determines when to stop selecting additional predictors. In QSAR, models with between 1 and 10 predictors are considered parsimonious and interpretable.

To verify the robustness of the subset with selected predictors, forward selection can be applied in a LOOCV context. Especially in smaller size datasets, a outlying observations can lead to the selection of different predictors. Performing the selection in a LOOCV allows investigation of the frequency by which predictors are selected. The predictors most frequently selected are used in the final model.

Forward selection can be applied to both linear and nonlinear additive models. Whereas it is frequently used in combination with MLR, the combination of forward selection with GAMs is less common. Since MLR models and GAMs are both additive, no problems are expected with the one-by-one addition of predictors and the assessment of the model fit. What should be kept in mind is that GAMs provide more flexibility. Addition of predictors that capture noise to a GAM might introduce overfitting, resulting in a low RSS. Such predictors can thus be selected as best predictor while this is not desired.

2.8 Model validation

The assessment of QSAR model performance is important because accurate predictions on the activity of new molecules can only be trusted when the model fits well and is reliable. The performance of QSAR methods can be assessed through internal and external validation methods. This thesis is limited to internal validation methods as

no sufficient data is available for separation into training and test sets. The internal validation methods used are the significance of the models and their individual predictors ($p < 0.05$), coefficient of determination R^2, $R^2_{adjusted}$ which is corrected for the number of descriptors used in the model, and the leave-one-out cross validated coefficient of determination Q^2_{LOO}.

The coefficient of determination R^2 quantifies the proportion of variance in the outcome that a statistical model can explain using the selected predictors. R^2 ranges from zero to one, where a value close to one indicates more variation explained by the model and thus a better fit. The R^2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2.}$$

R^2 is only valid in linear context and tends to increase upon the addition of every predictor, even if their inclusion does not significantly improve the model performance. This hinders comparison of models with different numbers of predictors and can lead to misleading conclusions. These limitations are addressed by the adjusted version of R^2, denoted as $R^2_{adjusted}$. $R^2_{adjusted}$ corrects for the number and usefulness of predictors included. Mathematically, $R^2_{adjusted}$ is determined as:

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(n - 1)}{(n - p - 1)},$$

where p is the effective degrees of freedom of the model and n the number of observations in the data. In a linear context, p equals the number of predictors in the model (Fox, 2015). In GAMs, the correction for effective degrees of freedom allows penalization for the flexibility and complexity introduced by the smoothing splines. Therefore, $R^2_{adjusted}$ is also valid in that context.

R^2 and $R^2_{adjusted}$ only explain how well the model fits the data, yet another important aspect of a model is its ability to predict the outcome for new observations. The leave-one-out cross validated coefficient of determination Q^2_{LOO} is a measure that quantifies the internal predictive power of a model. This parameter allows for comparison of different QSAR models in a simple way by presenting the values in a standardized range. The values for Q^2_{LOO} range from zero to one but can dive below zero when a model's internal predictive ability is very bad. The latter suggests modelling issues like overfitting.

To calculate Q^2_{LOO}, models are fitted in a LOOCV context. The model fitted on all molecules in the data minus one is used to predict the outcome of the molecule that was left out. Each molecule is left out once and thereby a prediction on the outcome is obtained for each of them. Since these molecules are extracted from the data, their observed value is also known. Using the predicted and observed values from each molecule, the value of Q^2_{LOO} is calculated as:

$$Q^2_{LOO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

with n the number of molecules (Consonni et al., 2009; Todeschini et al., 2009). This equation looks identical to the equation for R^2, but there predictions are made using the model fitted on all molecules.

According to Tropsha (2010), a QSAR model is predictive when $Q^2_{LOO} > 0.5$ and $R^2 > 0.6$. QSAR models satisfying these conditions are expected to fit the underlying patterns in the data. Models with high R^2 and low Q^2_{LOO} likely overfit the data. Such models explain a large part of the variance in the outcome by capturing noise. Due to this overfitting, these models cannot generalize patterns in the data and their predictive ability is low.

2.9 The Applicability Domain

A very important concept in QSAR is the Applicability Domain (AD). The AD defines the chemical space for a model. Predictions on the activity of molecules that do not belong to the same AD as where the model is trained on are invalid. Also, not all molecules in a dataset might belong to the same AD. Determination of the AD by robust statistical methodology prevents extrapolation errors and identifies the boundaries of a model's reliability (Todeschini et al., 2009).

The identification of the chemical space occurs in two ways. It is important to make a well-considered choice from a chemical perspective. One should choose a group of molecules that is expected to belong to the same AD. In this thesis, molecules from the class of prenylated phenolics are tested. Prenylated phenolics have similar basis skeletons but differing side groups ensure division into different subclasses (see figure 1). The homogeneity of these subclasses should be assessed. Here, the Chalcones subclass is very different from the other prenylated phenolic subgroups as they only have 2 ring structures instead of 3. This structural difference can cause a different mode of action for the Chalcones which may not align with the other prenylated phenolics. The influence of deviating subclass structures within a molecule class is assessed by fitting models with and without this subclass. A strong decrease in model performance is in an indication that the subclass does not belong to the same AD.

Next to these intuitive considerations, the fit of the individual molecules to the AD should be investigated. No standard protocol is available for this yet (Tropsha, 2010). In linear context, the leverage approach is commonly used. Leverage is a similarity measure that quantifies how far away a molecule is from the other molecules in the model with respect to its predictor values. A high-leverage point is considered an outlier as it has no close neighbors in the space defined by the predictors. Resultingly, they can cause drastic changes in the estimated model coefficients upon their removal. Even though high-leverage points are often also high influence points, this is not necessarily the case.

Using the predictors selected by a linear model, the leverage or hat matrix is obtained by:

$$H = X(X^T X)^{-1} X^T,$$

where the diagonal elements represent the leverages for each of the molecules. The leverage threshold is defined as three times the number of predictors divided by the number of molecules the model is fitted on. Molecules exceeding this threshold are considered high leverage (Sahigara et al., 2012).

The molecular descriptors in this study are selected using LOOCV. Therefore, the leverages are also determined in a LOOCV context. The number of leverages obtained for each molecule equals the total number of molecules minus one as each molecule is left out of the data once. For all resulting LOOCV models, high-leverage molecules can be identified. The molecules most frequently marked as high leverage are excluded from the data to assess their influence on the model performance.

Visualization of leverages is done using Williamsplots. The leverages are plotted on the horizontal axis and the standardized residuals on the vertical axis. An example of a Williamsplot with the leverage threshold indicated in red is shown in figure 6.

In nonlinear regression, determination of the AD needs additional attention. Very limited studies on concepts comparable to leverage in a nonlinear context are available. This thesis proposes a method based on internal predictive power Q^2_{LOO} to assess how the presence of each individual molecule influences the model performance. Fitting the model with an additional LOOCV around it allows to compute the Q^2_{LOO} on datasets where each molecule is ignored once. The influence of the molecule that is left out can be assessed by comparing the Q^2_{LOO} from the corresponding model with the Q^2_{LOO} from

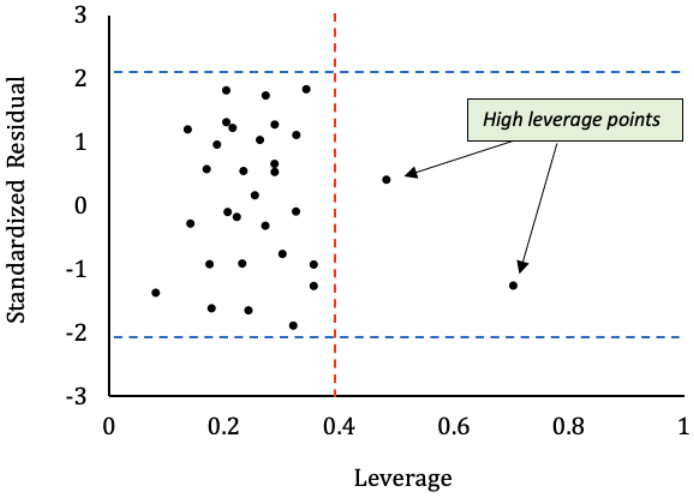

Figure 6: Williamsplot with the leverage threshold indicated with the red dotted line. The blue dotted lines represent the residual thresholds.

the model fitted on all molecules. If the removal of this molecule leads to an increase in Q^2_{LOO}, its removal benefits model performance. This molecule can then be said to fall outside the AD.

3 Results

3.1 Forward-MLR

Forward-MLR models with the number of predictors ranging from 1 to 10 were fitted on different compositions of the experimental dataset. Hereby, the influence of imputed MIC values for inactive prenylated phenolics and the inclusion of the Chalcones subclass on the fit ($R^2_{adjusted}$) and internal predictive power (Q^2_{LOO}) of the models was investigated. The addition of imputed values for the inactive compounds nor the addition of the Chalcones subclass improved the fit and the internal predictive power for the MLR models. Resultingly, the best experimental dataset consisted of 28 prenylated phenolics that were tested against the same strain of *S. mutans*.

This experimental dataset was complemented with the literature dataset that allowed for different strains of *S. mutans* and the other literature dataset that allowed for different *Streptococcus* species. Neither of the literature additions improved the model in terms of fit or internal predictive power. Therefore, the best dataset remained the experimental set with 28 prenylated phenolics.

None of the forward-MLR models fitted on this best experimental data were statistically compliant ($Q^2_{LOO} > 0.5$ and $R^2 > 0.6$). Figure 7 shows that the highest internal predictive power was obtained for the two-predictor MLR model (Q^2_{LOO}= 0.358).

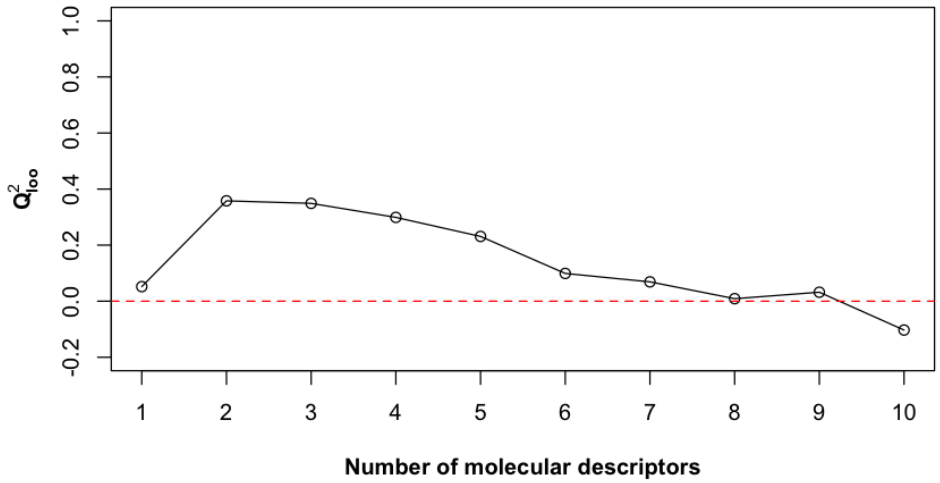

Figure 7: The internal predictive power (Q^2_{LOO}) of Forward-MLR models with 1-10 molecular descriptors fitted on the experimental dataset with 28 prenylated phenolics. The red dotted line indicates $Q^2_{LOO} = 0$.

The frequencies by which the LOOCV selected molecular descriptors for the two predictors are shown in the bar plots in figure 8. For the first predictor, the first atomic-level cut in the Partial Equalization of Orbital Electronegativity charge calculation (*GCUT_PEOE_1*) was selected in 82% of the models. Polar volume at -2.0 kcal/mol (*vsurf_Wp4*) was selected as the second predictor in 82% of the models. The two-predictor MLR with these selected predictors was fitted on the dataset with 28 prenylated phenolics. The resulting model equation for this model was:

$$pMIC = 6.815(0.419) + 5.372(0.942) * GCUT\_PEOE\_1 – 0.032(0.010) * vsurf\_Wp4,$$

where *GCUT_PEOE_1* is positively correlated with the pMIC and *vsurf_Wp4* negatively. The model was significant ($p = 6e{-}06$) even as the two molecular descriptors ($p = 6e{-}06$ and $p = 3.e - 03$, respectively). The $R^2_{adjusted}$ corresponding to this model was 0.586.

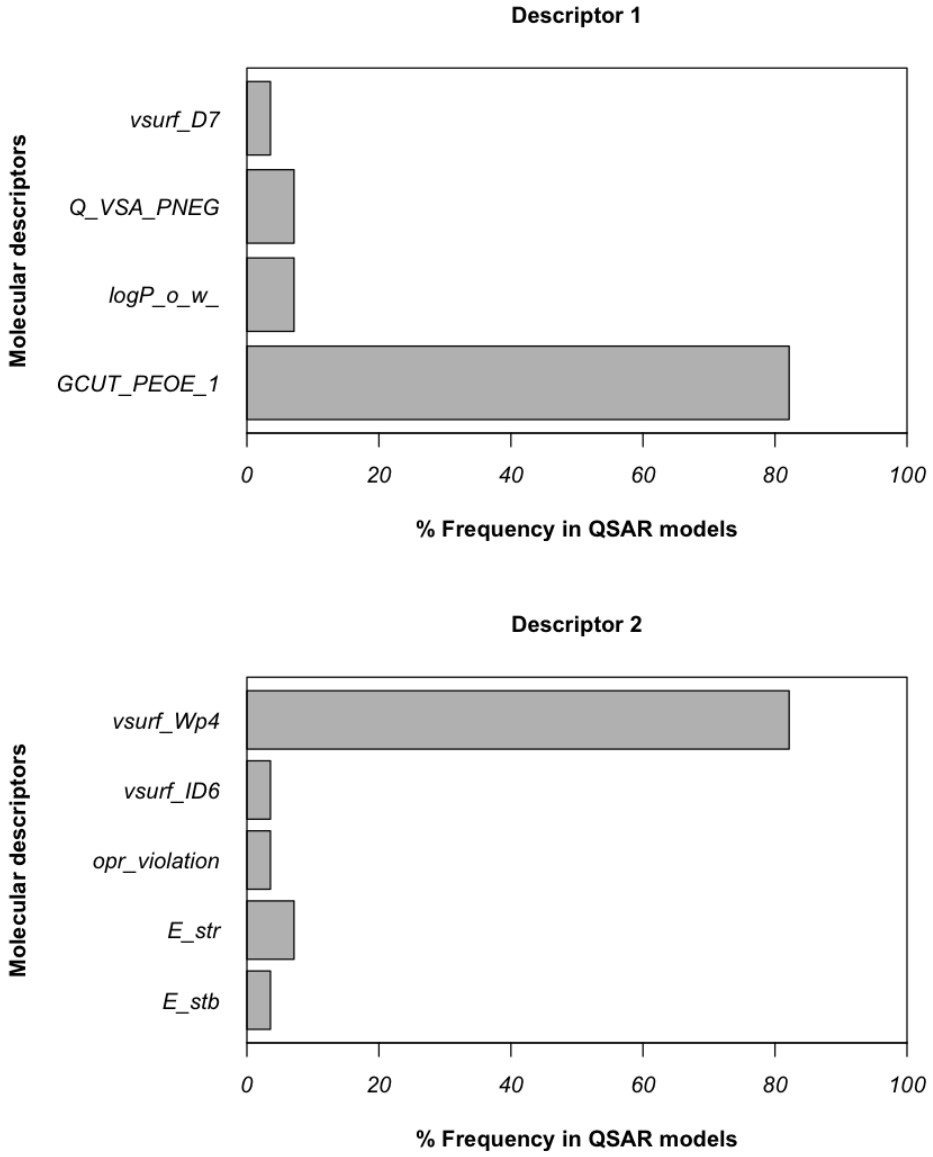

Figure 8: Bar plots showing the most frequently selected molecular descriptors for the two-predictor MLR model fitted on the experimental dataset with 28 prenylated phenolics.

The AD for this two-predictor MLR model was investigated. Figure 9 shows a bar plot with the frequencies by which prenylated phenolics were selected as high leverage molecules in the LOOCV. In more than 80% of the models, *Isokanzonol V* and *Glyasperin C* were selected. Therefore, they were suspected of not belonging to the same AD as the other prenylated phenolics. To check if the final two-predictor MLR model selects the same high leverage molecules as the LOOCV, a Williamsplot was made as shown in figure 10. From this plot can be seen that indeed *Isokanzonol V* and *Glyasperin C* had a leverage above the threshold indicated with the dotted red line.

Removal of the two high leverage molecules resulted in decreased internal predictive power (Q^2_{LOO} = - 0.208) and a decreased fit ($R^2_{adjusted} = 0.536$). The low Q^2_{LOO} combined with a relatively high $R^2_{adjusted}$ suggests overfitting. Resultingly, the two-predictor MLR model fitted best on all 28 prenylated phenolics.

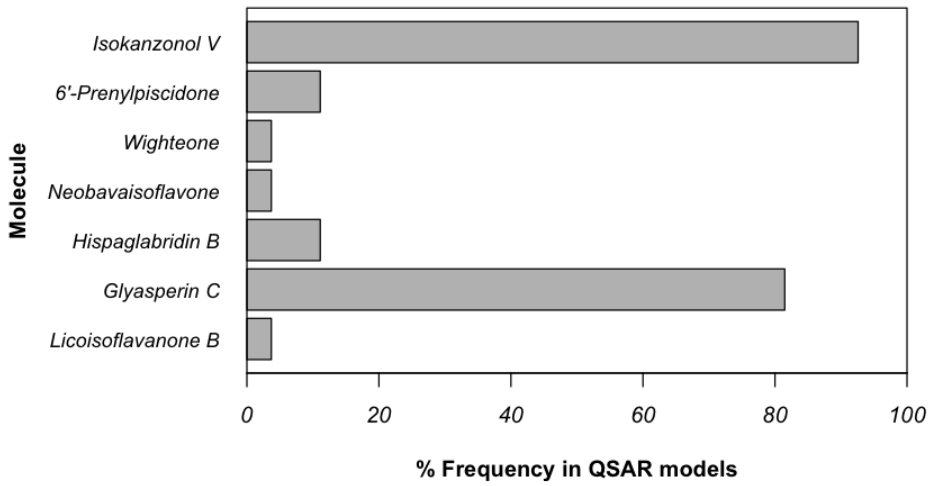

Figure 9: Bar plot showing the frequency by which molecules were selected as high leverage molecules in the LOOCV for the two-predictor MLR model fitted on the experimental dataset with 28 prenylated phenolics.

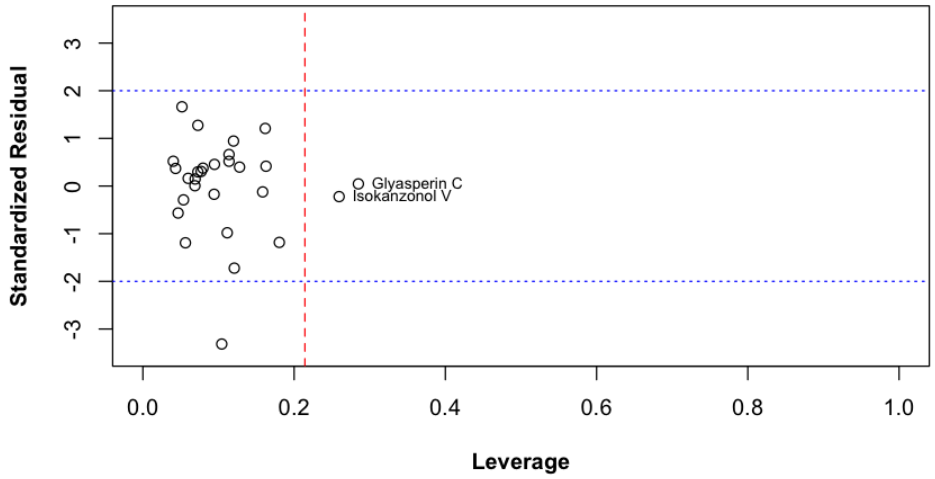

Figure 10: Williamsplot for the two-predictor MLR model fitted on the experimental dataset with 28 prenylated phenolics.

Figure 11 shows the pMIC values predicted from this model versus the observed pMIC values. Prenylated phenolics with a difference of more than 0.2 M between their predicted and observed pMIC are labelled. From this figure can be seen that seven molecules are labelled, which corresponds to the low predictive ability of this model.

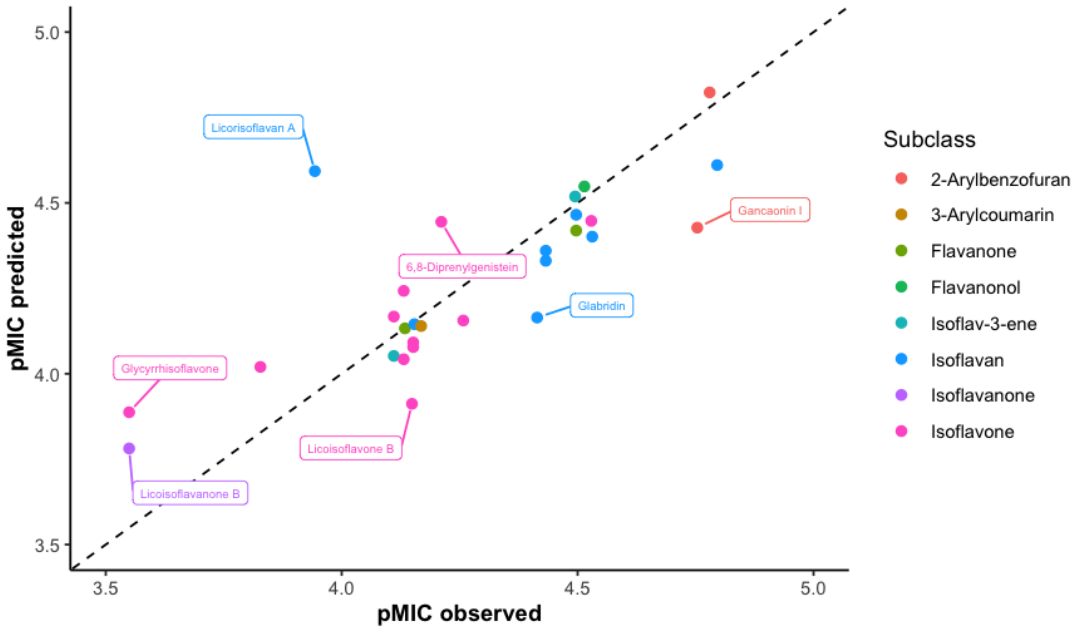

Figure 11: pMIC predicted versus pMIC observed for the two-descriptor MLR model on the experimental dataset with 28 prenylated phenolics. Prenylated phenolics with a difference $>$ 0.2 M between the observed and predicted pMIC are labelled. The prenylated phenolics are coloured by subclass.

3.2 Forward-GAM with smoothing splines

Forward-GAMs with molecular predictors modelled as smooth terms were fitted with 1-10 molecular descriptors and 3-10 basis functions. These forward-GAMs were fitted on different compositions of the experimental dataset to investigate the influence of imputed MIC values for inactive prenylated phenolics and the inclusion of the Chalcones subclass on the fit ($R^2_{adjusted}$) and internal predictive power (Q^2_{LOO}). The number of predictors in these datasets was reduced following the constraint that the number of unique values for each predictor should exceed the number of basis functions used in the smoothing spline. Therefore, all predictors with less than 10 unique values were removed.

The addition of imputed MIC values for the inactive compounds nor the addition of the Chalcones subclass improved the fit and internal predictive power for the forward-GAMs. For the resulting experimental dataset with 28 prenylated phenolics, the heatmap with the Q^2_{LOO} values for each of the fitted models is shown in Figure 12. It was observed that all models with one predictor and two models with two predictors showed a positive Q^2_{LOO}. Also, the model with four predictors and seven basis functions showed a positive Q^2_{LOO}. GAMs with an increased number of predictors suggested overfitting as their internal predictive power went below zero while the corresponding $R^2_{adjusted}$ values were exceeding 0.6.

The resulting experimental dataset was complemented with the first literature dataset that allowed for different strains of *S. mutans*. This addition caused the internal predictive power for all the models to go below zero. The dataset was further augmented with the literature data that allowed for different *Streptococcus* species. This addition increased the internal predictive power to positive values for models with 1 - 6 predictors and 3 - 5 basis functions, but none of them was statistical compliant. Therefore, the experimental dataset with 28 prenylated phenolics remained the best dataset for fitting forward-GAMs.

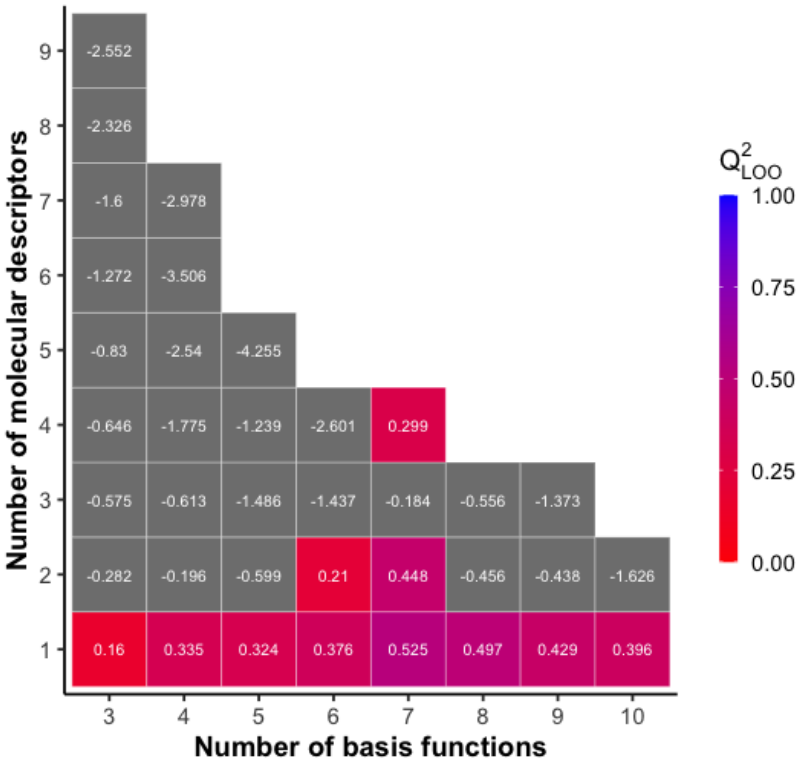

Figure 12: Heatmap with the internal predictive power (Q^2_{LOO}) of Forward-GAMs with 1-10 molecular descriptors and 3-10 basis functions on the experimental dataset with 28 prenylated phenolics.

As was observed from the heatmap in figure 12, the forward-GAM with the best internal predictive power (Q^2_{LOO}= 0.525) had one predictor and seven basis functions. Hydrophobic volume at an interaction energy of -0.2 kcal/mol (*vsurf_D1*) was selected as the best molecular descriptor in 100% of the LOOCV models. The model with *vsurf_D1* as predictor and seven basis functions was fitted. Both the model and *vsurf_D1* were significant ($p = 2e-16$ and p = $7e-5$ respectively). Since the corresponding $R^2_{adjusted}$ was 0.638, this model was statistically compliant and well performing.

For this best GAM, the AD was investigated. The internal predictive power calculated by the additional LOOCV around the original internal predictive power computation was used to assess the influence of individual prenylated phenolics on the model performance. The removal of *Glycyrrhisoflavone*, *Wighteone*, and *Licoisoflavone A* resulted in Q^2_{LOO} values of respectively 0.582, 0.575 and 0.547. Their removal thus increased the internal predictive power compared to the GAM that was fitted on all 28 prenylated phenolic. Simultaneous removal of *Glycyrrhisoflavone*, *Wighteone*, and *Licoisoflavone A* resulted in improved internal predictive power (Q^2_{LOO}= 0.607) and improved model fit (R2adj = 0.733). The smooth term of *vsurf_D1* was significant ($p = 3e-16$) even as the model itself ($p = 2e-16$).

Since the model equation for a GAM with smooth terms is not as interpretable as in MLR, the final forward-GAM was interpreted using the profile plot shown in figure 13. Here, the pMIC values as predicted by the smooth term are plotted against the molecular descriptor *vsurf_D1*. The smoothness of the curve was quantified by 5.52 EDF. The little bars on the x-axis show the values of *vsurf_D1* that were represented in the dataset. Compared to other molecular descriptors in the data, the observed values for *vsurf_D1* were relatively well spread across its range. Inspection of the profile plot

shows that the highest pMIC values are expected when the value of *vsurf_D1* is between 930 and 1050 kcal/mol.

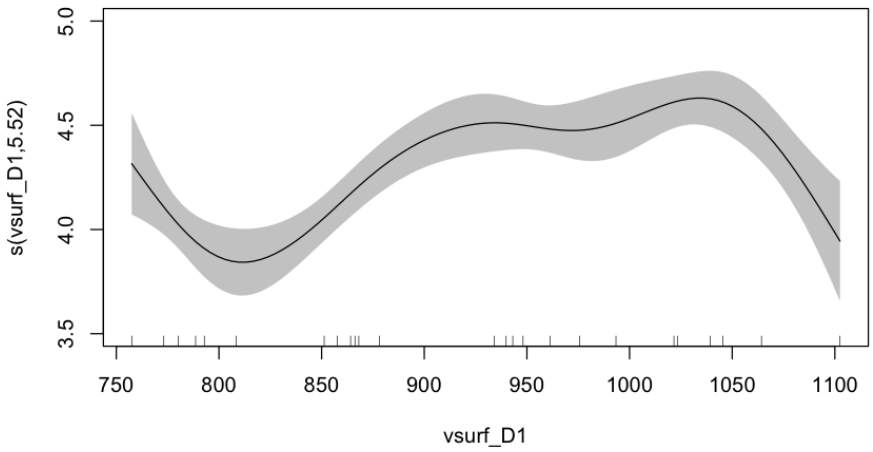

Figure 13: Profile plot for the molecular descriptor *vsurf_D1* that was fitted with seven basis functions on the experimental dataset with 25 prenylated phenolics.

The plot with predicted pMIC values versus observed pMIC values for the final forward-GAM is shown in figure 14 . Outliers with a deviation of more than 0.2 M between their predicted pMIC and observed pMIC are labelled. Most of the points were observed closely around to the black dotted line, which reflects the statistical compliant internal predictive power of the model.

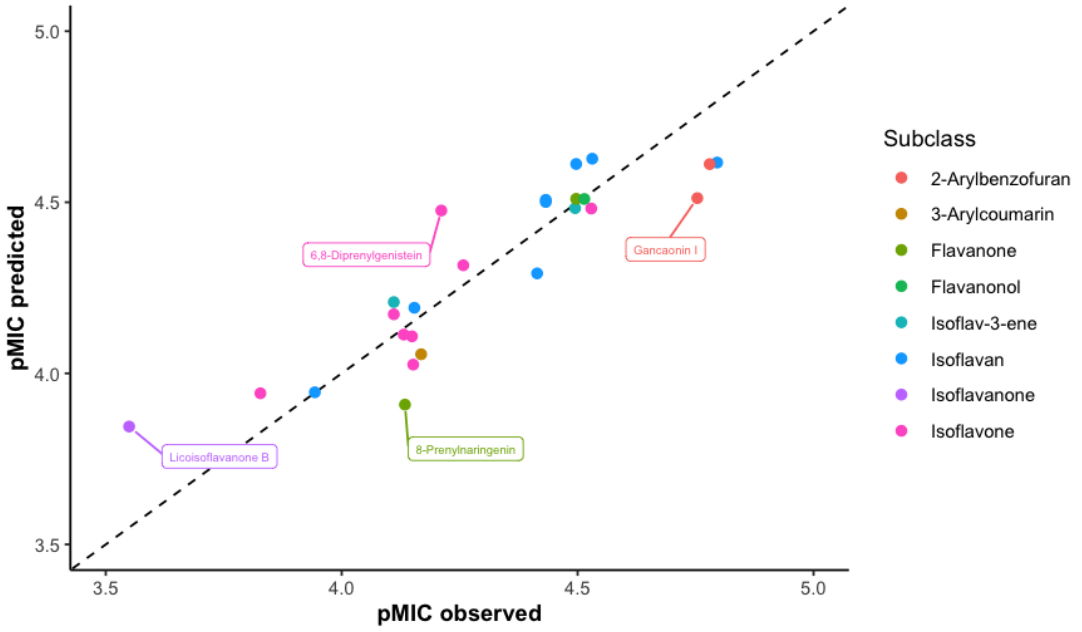

Figure 14: pMIC predicted versus pMIC observed for the one-descriptor forward-GAM with seven basis functions fitted on the experimental dataset with 25 prenylated phenolics. Prenylated phenolics with a difference > 0.2 M between the observed and predicted pMIC are labelled. The prenylated phenolics are coloured by subclass.

4 Discussion

By fitting forward-MLR models and forward-GAMs with smooth terms on different compositions of the data, the effect of imputed values for the MIC of inactive molecules, addition of the Chalcones subclass, and literature data was assessed. The imputation of MIC values for molecules that did not inactivate *S. mutans* at the highest concentration tested caused a decrease in model performance for both forward-MLR models and forward-GAMs. Because the imputations can be seen as a calculated guess, they introduced a lot of uncertainty. The observed decrease in model performance suggests that imputated MIC values were too uncertain in this case.

The addition of the Chalcones subgroup also led to a decrease in model performance for the forward-MLR models and forward-GAMs. This observation was in accordance with the hypothesis that the remarkable difference in ring structures between Chalcones and the other prenylated phenolics in the dataset would lead to differences in their structural features and mode of action, causing them to belong to a different AD. It should be recognized that the subgroups are not proportionately represented. The Chalcones represent 14% of the experimentally tested molecules, and some subclasses are represented by only one molecule. When aiming to ensure the fit of every subclass to the same chemical space, it is recommended to gather data more balanced with regards to subclasses. This allows to leave each subclass out of the model once and assess their influence on the model performance.

Incorporation of literature data did also not improve model performance in both forward-MLR models and forward-GAMs. This suggests that the level of noise introduced by allowing for different bacterial strains and species is too large to be captured by these models. Additional noise is included by the different experimental procedures that was allowed for, despite articles were selected to resemble the experimental conditions used by the in-house experiments as much as possible. The inherent heterogeneity in the data introduces complexities that challenge the model's ability to discern meaningful patterns. When considering the addition of literature data, the quality and compatibility should be carefully considered to avoid the introduction of too much noise.

On the resulting best experimental dataset with 28 prenylated phenolics, the performance of forward-MLR models and forward-GAMs was further investigated. The performance of all forward-MLR models was below the critical boundaries ($Q^2_{LOO} > 0.5$ and $R^2 > 0.6$), whereas GAMs with smooth terms for the predictors showed values close to or exceeding these boundaries. The linearity assumption in MLR forces the forward selection procedure to select predictors that best fit a linear relation with the outcome. These predictors do not necessarily explain a large part of the variation in the outcome. Other predictors can explain more variation, but only when the model allows them to follow their relation to the outcome. This explains why smooth terms increased the model performance and emphasizes the importance of inspecting the functional form of the relation between the predictors and the outcome before choosing a modelling method.

Resulting from the small size of only 28 prenylated phenolics of the final dataset, multiple predictors in the dataset showed unequal spread of values across their range. Both MLR and smoothing splines are affected by this, but the extend of the impact varies. In MLR, unequal spreads of the predictor values may affect the estimates of the coefficients and its model performance, but due to the linearity assumptions no underlying relations need to be interpolated. The latter is what causes a challenge in smoothing splines. Here, the underlying trend on intervals with sparse or no data becomes difficult to capture. The interpolation of the curve in these gaps can lead to biased trends. With regards to the dataset, it is therefore important to have enough observations such that the spread of predictors values is as uniform as possible.

Not only the spread of predictor values, but also the spread of the outcome, pMIC, in this dataset was problematic. All prenylated phenolics tested had a pMIC value between 3.5 and 4.8. This limited variation causes difficulties in identifying the underlying relationship which results in uncertain and biased parameter estimates. Also, models build on data with limited outcome variation struggle generalize to other datasets. This poses challenges in the model validation performed using LOOCV where the yielded estimates may thus not be reliable. The lack of variation for the pMIC is partly caused by the fact that MIC values were only measured up to a concentration of 100 µg/mL which results in a lower bound for the pMIC values. It is recommended to change the experimental design such that a broader range of pMIC values is obtained.

The AD in a linear context was assessed in terms of leverage. Removal of the identified high leverage molecules *Isokanzonol V* and *Glyasperin C* from the best two-predictor forward-MLR model did not improve but even decrease model performance. High leverage molecules have extreme predictor values or unusual combinations of predictor values compared to the other observations present in the data. According to Alguraibawi et al. (2015) there exist two types of high leverage points. Good high leverage points have extreme predictor values, but their outcomes follow the data pattern. Therefore, they contain valuable information about the extreme regions of the predictors. These points provide stabilization and should not be removed from the data. Bad high leverage points deviate from the data pattern causing strong changes in the estimates of the regression coefficients.

Both *Isokanzonol V* and *Glyasperin C* did show extreme values for the molecular descriptors *vsurf_Wp4* and *GCUT_PEOE_1* on which their leverage was assessed, but their pMIC values did not deviate far from the estimated regression line. Also, both molecules had values at the border of the *vsurf_Wp4* range, relatively far away from the *vsurf_Wp4* values of the other molecules. This indicates that *Isokanzonol V* and *Glyasperin C* are good high leverage molecules with a stabilizing character. Especially in smaller datasets as in this study, the removal of good high leverage points causes a large fraction of information to be lost. This explains why the removal of *Isokanzonol V* and *Glyasperin C* did not improve the model performance for the best two-predictor forward-MLR model.

For the best GAM with *vsurf_D1* modelled as a smooth term, molecules with a negative influence on the internal predictive power (Q^2_{LOO}) were identified. The top three molecules showing the largest improvement were removed from the dataset. This resulted in an increase of the model performance. The proposed method seems to be well-defined and able to identify the AD in a nonlinear context. Since it measures the influence on Q^2_{LOO}, there is no risk of accidentally removing a molecule that has a stabilizing effect. This would namely cause a decrease in Q^2_{LOO}. Although this method performs well, it has the drawback of being computationally very intensive as a LOOCV is applied around another LOOCV. Computing times rapidly increase for larger datasets, leading to a computational burden when aiming to determine the AD for multiple different models or datasets.

How the forward-GAMs with smooth terms for the predictors were fitted in this thesis can be discussed on several aspects. Usually, the upper limit of k for a smooth term is determined by the number of unique datapoints a predictor has. In this thesis, the smooth terms in the forward-GAMs were assigned fixed numbers of basis functions k ranging from 3 to 10 to explore the effect of restricting wigglyness. Therefore, predictors with less than 10 unique values were removed from the dataset prior to fitting a GAM. Previous studies showed that the variables removed by this restriction are not commonly selected as best predictors. Yet, this should not be generalized to other studies as in QSAR there is often no a priori knowledge on what molecular descriptors are related

to the outcome (Todeschini et al., 2009). Additionally, the majority of those studies used linear models and therefore certainty is lacking on the importance of the removed predictors in nonlinear context. To prevent ignorance of important predictors, the removal of predictors should rather be limited. This approach requires more flexibility in k, as continuous predictors with less unique datapoints can only be modelled with lower values of k.

This more flexible manner of using k can further benefit smooth terms in forward-GAMs. Where λ is optimized by Generalized Cross-Validation to balance smoothness and fit, k also plays a role in the optimization of this balance. In this study, the smooth term for the best forward-GAM was modelled using seven basis dimensions. The number of basis functions is related to the upper limit of degrees of freedom for the smooth term as k–1 = 6 degrees of freedom. It was found that the EDF for this smooth term (5.518) closely approximates the 6 degrees of freedom. EDF close to the upper limit of degrees of freedom indicates an overly constrained model in terms of wigglyness. Therefrom, it follows that the balance between smoothness and fit of this model is not yet optimal. Optimization of k for each individual predictor instead of using a range of fixed k's allows the model to further optimize this balance and gain improved performance.

5 Future recommendations

This thesis focused on the basic principles and applications of smoothing splines to explore its usefulness in QSAR modeling. To enhance the performance of QSAR models even further, it is recommended to examine the trends captured by smoothing splines and investigate if they can be represented by more accurate modelling manners. An example is a smoothing spline that shows a threshold value above which the activity is increased. This suggests the use of a step function to capture the relation more precise. Using GAMs, each of the predictors can be fitted by the type of function that is most effective for them.

To our knowledge, this thesis is one of the first to extensively describe the use of smoothing splines in QSAR. For a comprehensive evaluation on the functionality and generalizability of smoothing splines, it is advised to test its performance on diverse datasets with distinct characteristics. The dataset used in this study was small and exhibited limited variation in outcome, with predictor values not always well spread over their range. Training on diverse datasets with more optimal characteristics can verify if the usefulness of smoothing splines.

Where the proof of the pudding lies in the eating, the proof of QSAR models lies in their external validation. Models that perform excellent in internal validation, can still completely fail external validation. Insufficient results from internal validation however always indicate unreliable models. It is therefore highly recommended to apply external validation before relying on predictions from an internally well-predicting model. Hereby, it is of great importance that the molecules used for the external validation fall in the same AD as the molecules the model is trained on (Consonni et al., 2009; Golbraikh & Tropsha, 2002).

Investigation of the AD by a robust method is very important for the reliability of predictions. Robust AD methodologies provide a safeguard against extrapolation errors, identifying the boundaries of the model's reliability. By addressing the question of which molecules to include or exclude from the modelling, QSAR models can reach higher accuracy, giving more confidence to the use of the model. The leverage approach used in linear context requires a manual check to distinguish between good and bad high leverage points. Here, it is advised to investigate different measures that do not require manual inspection afterwards. The proposed method to determine the AD in nonlinear context needs further verification and optimization to overcome the computational burden.

Data collection in QSAR studies often results in incomplete or censored observations. The exact activity is not measured, only that it must be above a certain threshold. Common methods to deal with incomplete observations like row-wise deletion inefficiently discard data and data imputation methods might be too crude. Censored regression approaches maximize the use of available information without introducing potentially inaccurate imputed values, providing a more robust method for handling incomplete data in QSAR studies.

Lastly, QSAR models face hurdles like the need for a certain amount of data but also have a limited applicability outside the domain the model was trained for. Even with sufficient data, these models serve specific tasks. For new tasks (e.g. prediction a MIC for a different bacterial species) the activities for all molecules need to be (re-)measured and remodelled. To tackle this, a transfer learning (TL) approach is recommended. TL leverages knowledge gained in one scenario to enhance performance in a different situation. Essentially, it allows the reuse of existing QSAR models, minimizing the need for new data collection and making the process more efficient. This approach ensures that valuable insights gained from one domain can be applied to another, speeding up research and potentially saving time and resources.

6 Conclusion

This study proposed a nonlinear QSAR modelling method using forward-GAMs with the predictors modelled as smooth terms. Different compositions of the dataset were assessed, and concluded was that the experimental data without imputed MIC values for inactive prenylated phenolics and the Chalcones subgroup produced the best models. This dataset contained a total of 28 prenylated phenolics. The forward-GAM with one predictor modelled as a smoothing spline was statistically compliant while the baseline MLR models were not. The best predictor selected by this best forward-GAM was hydrophobic volume at an interaction energy of -0.2 kcal/mol (*vsurf_D1*). Visualization by means of profile plots ensured interpretability for the GAMs. The AD was determined in a linear and nonlinear context to investigate how this enhances QSAR modelling. In the linear context, the leverage approach identified two high leverage molecules. Their removal did not enhance the forward-MLR model performance. Further inspection concluded that these molecules were good high leverage molecules and had a stabilizing effect on the MLR. For the nonlinear forward-GAMs, the AD was investigated by assessing the removal of molecules on the internal predictive power. The top three molecules whose removal most benefited the internal predictive power were removed which improved the fit and internal predictive power of the best forward-GAM even further. Overall, it can be concluded that the use of smoothing splines enhances QSAR model performance by allowing for more flexibility.

7 Acknowledgements

I would like to thank my supervisor Jos Hageman for his enthusiastic guidance throughout this thesis. We had insightful discussions that spiked my interest in QSAR. I would also like to thank Sarah van Dinteren, Janniek Ritsema and Carla Araya-Cloutier for sharing the results and findings of their extensive labwork regarding the anti-*Streptococcus* activity of prenylated phenolics.

8 References

Alguraibawi, M., Midi, H., & Imon, A. H. M. R. (2015). A New Robust Diagnostic Plot for Classifying Good and Bad High Leverage Points in a Multiple Linear Regression Model. Mathematical Problems in Engineering, 2015, 1–12.

Balasundram, N., Sundram, K., & Samman, S. (2006). Phenolic compounds in plants and agri-industrial by-products: Antioxidant activity, occurrence, and potential uses. Food Chemistry, 99(1), 191–203.

Chang, S. K., Jiang, Y., & Yang, B. (2021). An update of prenylated phenolics: Food sources, chemistry and health benefits. In Trends in Food Science and Technology (Vol. 108, pp. 197–213). Elsevier Ltd.

Consonni, V., Ballabio, D., & Todeschini, R. (2009). Comments on the definition of the Q2 parameter for QSAR validation. Journal of Chemical Information and Modeling, 49(7), 1669–1678.

Fox, J. (2015). Applied regression analysis and generalized linear models.

Golbraikh, A., & Tropsha, A. (2002). Beware of q2! Journal of Molecular Graphics and Modelling, 20(4), 269–276.

Hageman, J. (2022). Relevant metabolites' selection strategies. In Metabolomics Perspectives: From Theory to Practical Application (pp. 381–398). Elsevier.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning (Second Edition, pp. 289–327). Springer .

Kowalska-Krochmal, B., & Dudek-Wicher, R. (2021). The minimum inhibitory concentration of antibiotics: Methods, interpretation, clinical relevance. In Pathogens (Vol. 10, Issue 2, pp. 1–21). MDPI AG.

Liao, Y., Brandt, B. W., Li, J., Crielaard, W., Van Loveren, C., & Deng, D. M. (2017). Fluoride resistance in Streptococcus mutans: a mini review. In Journal of Oral Microbiology (Vol. 9, Issue 1). Taylor and Francis Ltd.

Liu, P., & Long, W. (2009). Current mathematical methods used in QSAR/QSPR studies. In International Journal of Molecular Sciences (Vol. 10, Issue 5, pp. 1978–1998).

Murray, C. J., Ikuta, K. S., Sharara, F., Swetschinski, L., Robles Aguilar, G., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E., Johnson, S. C., Browne, A. J., Chipeta, M. G., Fell, F., Hackett, S., Haines-Woodhouse, G., Kashef Hamadani, B. H., Kumaran, E. A. P., McManigal, B., ... Naghavi, M. (2022). Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. The Lancet, 399(10325), 629–655.

Roy, K., Kar, S., & Das, R. N. (2015). A Primer on QSAR/QSPR Modeling. Springer International Publishing.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. Molecules, 17(5), 4791–4810.

Todeschini, R., Consonni, V., & Gramatica, P. (2009). Chemometrics in QSAR. In Comprehensive Chemometrics (pp. 129–172). Elsevier.

Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation. Molecular Informatics, 29(6–7), 476–488.

Weaver, S., & Gleeson, M. P. (2008). The importance of the domain of applicability in QSAR modeling. Journal of Molecular Graphics and Modelling, 26(8), 1315–1326.

Wood, S. N. (2003). Thin plate regression splines. Journal of the Royal Statistical Society Series B: Statistical Methodology, 65(1), 95-114.

World Health Organization: WHO. (2021, November 17). Antimicrobial resistance.