Exploiting mAb structure characteristics for a directed QbD implementation in early process development

Karlberg, Micael, von Stosch, Moritz; Glassey, Jarka

School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RN, United Kingdom.

Abstract

In today's biopharmaceutical industries the lead time to develop and produce a new monoclonal antibody takes years before it can be launched commercially. The reasons lie in the complexity of the monoclonal antibodies and the need for high product quality to ensure clinical safety which has a significant impact on the process development time. Frameworks such as Quality by Design are becoming widely used by the pharmaceutical industries as they introduce a systematic approach for building quality into the product. However, full implementation of Quality by Design has still not been achieved due to attrition mainly from limited risk assessment of product properties as well as the large number of process factors affecting product quality that needs to be investigated during the process development. This has introduced a need for better methods and tools that can be used for early risk assessment and predictions of critical product properties and process factors to enhance process development and reduce costs.

In this review we investigate how the Quantitative Structure-Activity Relationships framework can be applied to an existing process development framework such as Quality by Design in order to increase product understanding based on the protein structure of monoclonal antibodies. Compared to Quality by Design, where the effect of process parameters on the drug product are explored, Quantitative Structure-Activity Relationships gives a reversed perspective which investigates how the protein structure can affect the performance in different unit operations. This provides valuable information that can be used during the early process development of new drug products where limited process understanding is available.
Introduction

Monoclonal antibodies (mAbs) are therapeutic proteins that have gained increasing importance over the years mainly due to their clinical specificity and safety in treatments, but also because they can be applied to a wide spectrum of different ailments. Their increasing popularity can be seen in recent reports on the market for mAbs where the sales have increased from around $39 billion in 2008 to around $75 billion in 2013, making mAbs one of the fastest growing bioproduct groups [1]. This has led to many advances in improving the mAb manufacturing processes in areas such as process optimization [2, 3] and process control [4] to bring down manufacturing costs and time to market. However, the process development for manufacturing of mAbs is cost- and time-intensive due to the high product quality requirements that must be met to make the product clinically safe for commercial use. This is especially pronounced for the downstream operations where intensive characterisation of the unit operations (UOs) are needed to ensure high product quality and purity of the end product [5, 6].

Paradigms such as Quality by Design (QbD) have become an integral part of process development in today’s biotech industries with the goal of obtaining a high and more constant product quality with less risk of batch failures [7, 8]. The QbD paradigm provides a work-flow of building quality into the product by placing emphasis on increasing the process understanding and investigating how process parameters (PPs) affect the final product quality. The implementation of the framework has been covered extensively in many reviews [7, 9-11]. Continuous improvements are constantly being made to increase the effectiveness and applicability of these frameworks for production of biopharmaceuticals [12, 13].

However, effective implementation of QbD is still proving to be a significant challenge in biopharmaceutical industries due to the complex relationships between process parameters and product quality, which need extensive experimental studies to be characterized [14]. Eon-Duval et al. [15] summarized the most common Critical Quality Attributes (CQAs) for mAbs in process development that are important to monitor and control and how different PPs such as temperature, pH and ionic strength can have an effect on the product quality. An example of this is the glycan structure of the mAbs which is important for the efficacy and stability of the protein and is therefore an important factor to monitor and control [16-18]. In 2012, Genentech & Roche attempted to file a QbD application for the antibody Perjeta (pertuzumab) to Food and Drug Administration (FDA) but the application was rejected due to the design space not being properly characterized [19]. The reasons included lack of thorough risk assessment and the failure to recognize important PPs in the bioreaction of the mAb synthesis, which introduced high variability in the glycan structure. This had an undesired effect on the Antibody-Dependent Cell-mediated Cytotoxicity (ADCC), which made the antibody clinically unsafe. Thus, further characterization studies were needed to identify the critical process parameters (CPPs) and their optimal ranges to maintain desired glycan structure. Studies have shown that increased control over the glycan structure can be gained through optimizing the composition of the basal and feed media [20-22] or improving the cell line used for expression [23] which can help to drive the glycosylation towards the desired quality.

Sustained effort has gone into the development of more efficient high-throughput screening methods for both upstream and downstream operations to be able to perform more rigorous characterization studies. Many of these techniques are in small scale which helps to reduce the use of resources, costs and aids in speeding up process development [24], given that they are representative of the larger scale. This includes high-throughput screening of cell lines across different fed-batch scales [25], high-throughput development of media [26, 27], high-throughput process development using
Focus of this review

Process development of mAbs requires major investments from the pharmaceutical companies and they adopt the QbD paradigm because they expect that in the long term it will help to reduce the investments and reduce process development times. The idea is that a better understanding of the processes and products will allow reducing the experimental effort in the future. Today, the process and product knowledge is incorporated through risk assessment, specifying the product CQAs that need to be monitored and controlled within certain ranges, as well as listing process parameters that might have an impact on the specified CQAs. These assessments are usually based on experience from previous process developments of mAbs and literature research [33]. This is an iterative process, which requires a significant allocation of resources throughout process development and de novo for each new mAb.

Zurdo [34] suggested that the QbD framework needed to be extended to incorporate product knowledge in the form of developability of pharmaceuticals which includes manufacturability, safety/pharmacology and biological activity. The author argued that by using risk assessment tools based on in silico methods, information concerning different CQAs can be inferred by correlating structural data from many pharmaceuticals to their behaviour. In a later publication by Zurdo et al. [8], two case studies were presented where structural properties of the pharmaceuticals were linked to CQAs such as aggregation and half-life, which were successfully predicted.

Thus, the focus of this review explores the different aspects that need to be considered in order to produce in silico risk assessment tools for mAbs. In particular, the adoption of Quantitative Structure-Activity Relationship (QSAR) type modelling is reviewed as a method to incorporate mAb specific aspects into the QbD paradigm. It is argued that such tools can greatly aid risk assessment in early process development when not much is known about either the process or the product behaviour in different UOs of the process, e.g. propensity to aggregate in downstream processing. The means by which these methods can more efficiently reuse historical data from previous UO characterization studies of other mAb platforms in order to correlate the effect of generated structural features/properties and PPs on the product quality are also addressed. The methods of predicting the change of CQAs with different protein structures are discussed and it is argued that the optimal ranges can be inferred for PPs that affect the specified CQAs based on the historical process development data. It is concluded that in turn this should aid in reducing the number of experiments required to characterize the process, thus saving cost and development time.

Quantitative Structure-Activity Relationship (QSAR)

The QSAR framework relates structural features or descriptors of a compound to biological or physicochemical activity [35, 36]. This methodology was first introduced by Hammet [37] in the 1930s and was later refined by Hansch and Fujita and has become a standard tool for small drug discovery. A method derived from QSAR, referred to as Quantitative Sequence-Activity Modelling (QSAM), has been introduced in recent years and focuses on relating structural descriptors of proteins, peptides and
nucleic acids to activity [38]. Given the proteinaceous character of the mAbs, the QSAM methodology will be of more relevance and the workflow described below will therefore focus more on sequence based rather than small molecule based QSAR. The QSAR/QSAM framework have been applied to diverse range of challenges where structural properties of pharmaceuticals have been used directly for the prediction of different process related aspects such as the prediction of isotherm parameters in ion-exchange chromatography [39], ligand-binding in ion-exchange chromatography under high salt concentrations [40], binding of proteins in ion-exchange chromatography in different pH conditions [41], protein surface patch analysis for the choice of purification methods [42], chromatographic separation of target proteins from host cell proteins (HCP) [43], viscosity, clearance and stability prediction for antibodies [44] and degradation prediction of asparagine and aspartate in antibodies [45] to mention a few. This also showcases one of the main strengths of the QSAR/QSAM framework with its ability to link structural features to many different forms of prediction outputs. It is important to note, however, that identical experiments must have been performed on different pharmaceuticals in order to compare the differences in structure and their effect on the output. Equally important is that sufficient excitation is present in the output data in order for the effects to be linked to the corresponding structural feature.

A general outline of developing a QSAR model is presented in Figure 1. The structure of the outline is not representative for all types of data sets and will change depending on the area of application. It gives however, an overview of the vital components of the QSAR methodology which will be discussed further in the review.

![Figure 1. General outline of the QSAR model development. The represented structure is one of many possible layouts that can be used which depends on the type of the data, size of data set, number of descriptors, selection methods for descriptors, modelling method and goal of the model. However, the use of cross-validation and external testing should always be an integral part of all model construction to ensure high model quality and increase its ability to generalize when possible.](image-url)
**Descriptor generation**

One of the most important steps in QSAR/QSAM is how the structures of the pharmaceuticals in question can be described numerically in order to use them in correlation studies with prediction outputs of interest. For proteins such as mAbs two approaches to generate descriptors are discussed here: 1) descriptors generated from the amino acid primary sequence and 2) descriptors generated from three-dimensional models of the antibodies. It has been shown that a combination of both physicochemical and 3D descriptors works best and also ensures that the model is not overly reliant on a single type of a descriptor [46]. The workflow for the generation and incorporation of descriptors into QSAR is illustrated in Figure 2.

**Amino acid composition-based descriptor generation**

Extensive research has been carried out to develop new informative descriptors for peptides and proteins generated from their primary sequence [47]. This was first introduced by Sneath [48] who derived amino acid descriptors for the 20 naturally occurring amino acids from qualitative data. Later on Kidera et al. [49] used 188 properties of the 20 naturally occurring amino acids, which were converted into ten orthogonal new descriptors to describe the amino acids. Later the z-scale which consists of 3 new amino acid descriptors derived by applying principal component analysis (PCA) on 29 physicochemical properties [50, 51], was introduced. Other amino acid scales, which were also derived through PCA, include the extended Z-scale [52] and T-scale [53]. Other descriptors include the so called isotropic surface area (ISA) and the electronic charge index (ECI), which are derived from the three-dimensional structures of the amino acids [54]. All these descriptors were tested and performed well in respective studies on small peptides [51-54]. In a two-part review by van Westen et al [55, 56] many of the existing amino acid scales were benchmarked and compared. The authors demonstrated that the different scales described different physicochemical and topological properties which is useful when deciding on which scales to use. Obrezanova et al. [57] used several such amino acid scales to predict mAb aggregation propensity based on the primary structures. Douchinova et al. [58] applied the Z-scales descriptors to successfully predict ligand binding of peptides. However, even though amino acid descriptors explain differences in the primary structure, they do not take into consideration potential interaction between the amino acids in or between primary chains. It has been argued that this simplification can lead to a loss of information concerning properties of secondary and tertiary structure in larger proteins [47].

Descriptors can also be generated by using empirical equations on the entire primary sequence to infer protein properties such as the isoelectric point, hydrophobicity, molecular weight, physicochemical properties and secondary structure content to name a few. Many such tools and applications are available on bioinformatics sites, such as ExPASy [59] and EMBL-EBI [60].

**Homology modelling and molecular dynamics for descriptor generation**

Descriptors capturing structural and surfaces properties can be generated by using existing crystal or NMR structures or by building models using homology modelling. The latter is done by finding proteins with existing 3D structures that have a high level of similarity to the primary sequence of the protein of interest. These proteins are then used as templates to predict the likely structure of the queried protein [61]. This has been successfully used in many publications where information such as surface areas, angles and surface properties were extracted [43-45]. The method is especially useful when no crystal structure exists. Caution needs to be exercised, however, as the homology models are only predicted structures and might not represent the true protein conformation. Breneman et al. [62] introduced a methodology for generating 2D surface descriptors, also called transferable atom equivalent (TAE) descriptors, by reconstructing the electronic surface properties of the molecular structures from a library.
Figure 2. Examples of different routes for descriptor generation using the primary sequence and crystal structure of the protein. The primary sequence can be used to generate both local and global descriptors with the use of amino acid scales for numerical conversion and homology modelling/MD simulations to simulate surface properties. Available structure from crystal X-ray or NMR can be incorporated together with generated homology models. The different descriptors can be used together with modelling methods such as those mentioned in this review to link the structure of the protein to characterization data.
of atomic charge density components. This has the advantage of representing surface variations such as hydrophobicity and charge distributions numerically, which is of great importance when studying for example protein binding to an anion exchange chromatographic column packing using different salts [63]. Breneman et al. [64] later introduced the Property-Encoded Surface Translator (PEST) algorithm which is a further development to better describe the surfaces of the proteins when applying the TAE molecular surface descriptors. However, it is important to note that both TAE and PEST need 3D models in order to generate the descriptors of interest. PEST together with TAE descriptor, has been successfully applied in a QSAR study where the generated model was able to accurately predict protein separation from HCPs [43]. Robinson et al. [65] used the TAE descriptors to relate the structural differences between several Fab fragments to predict column performance between different chromatographic systems. It has been argued however that caution needs to be exercised when using library based descriptors as these are usually directly related to a specific state of a compound that was measured in a unique environment. This means that these descriptors should only be applied if experiments were carried out in an identical or similar environment. Otherwise, this might cause the descriptors to be biased [46]. Other structural properties, such as molecular angles and solvent accessible surface areas extracted from homology models, were used by Sydow et al. [45] to determine the risk of degradation of asparagine and aspartate in mAbs as post translation modifications. Similarly, Sharma et al. [44] investigated the risk of oxidation of surface accessible Tryptophans.

Due to the flexibility and size of the antibodies it is very difficult to produce good 3D structures based on X-ray crystallography and NMR. Instead, homology modelling has proven to be a good alternative to circumvent this problem. However, due to the size and the many flexible parts, such as loops, in the antibodies, pure homology models might not give a sufficiently accurate representation of the reality. Molecular dynamics (MD) is a useful tool that can be used to minimize the energy of the entire protein and to simulate the dynamics of the protein of interest in different environments [66]. MD simulations have also shown very high similarities in the internal dynamics of antibodies when comparing the simulated results to those observed in reality [67]. It can therefore be argued that MD simulation should be applied to all homology models before descriptors are generated to mimic the environment of the samples that are used in QSAR studies.

Table S1 shows a list of popular homology modelling and MD simulation software that has been used to generate 3D models of mAbs and extract descriptors.

Descriptor Selection

The descriptor generation usually results in a large number of descriptors being produced, requiring a descriptor subset selection such that the predictability of the QSAR model is increased by reducing noise and collinearity [68, 69]. This process is however very complex and despite extensive research, no clear workflows have yet been proposed. An established rule of thumb for building a good model is that the ratio between samples and descriptors should at least be five to one [68]. However, most studies struggle with a limited number of samples, which makes it hard to fulfil this condition. The choice of modelling method will have a great impact on this with methods like Support Vector Regression (SVR) and Partial Least Squares (PLS) being able to generate robust models even when the number of descriptors are greater than the number of samples [70, 71]. The performance of the selected descriptors is also dependent on the classifier that is used. An optimal combination is usually found through benchmarking different combinations of descriptor selection methods and modelling methods [72, 73]. It has also been shown that descriptor selection methods should not be purely data-driven but should also be based on prior knowledge and descriptors should be added if they are known to have an effect on the system [74].
The descriptor selection process can be separated into two parts; (1) dimensionality reduction before using the data in modelling and (2) model-based or algorithm-based descriptor selection when training the model, see Figure 1. In the first part, descriptor extraction and dimension reduction methods like PCA have proven to be an effective way to reduce collinearity in the initial descriptor set [75]. For the second part, many modelling methods have descriptor selection capability. For example, PLS weights the descriptor according to contribution to the model output, Support Vector Machines (SVM) select descriptors that can generate maximum distance between classes and Random Forest (RF) randomly selects descriptors to create unique trees for classification [76, 77]. Alternatively, there are algorithm-based selection methods such as forward selection or backward elimination where descriptors are added or eliminated from the model iteratively to reduce the modelling error and these are popular in QSAR studies [78]. For example, in the work of Chen et al. [79] the protein retention using hydrophobic interaction chromatography under different salt conditions was investigated. A sparse linear Support Vector Machines for Regression (SVR) algorithm was used for dimensionality reduction to select a subset of descriptors that had high correlation to the response before training a final nonlinear SVR model. For more extensive reviews on descriptor selection methods for QSAR refer to further literature, e.g. Shahlaei [68], Yousefinejad and Hemmateenejad [70] and Rudnicki et al. [80].

**Model Development and Validation**

*Cross validation*

The development of the model is carried out by training a classifier or regression method with a calibration data set that traditionally consists of 70-80% of all the available samples which can be seen in Figure 1. This should always be carried out with a resampling method in the form of cross-validation (CV) to fully make use of the available samples to optimize model parameters. The main advantage of using resampling methods is the increase in confidence in the final model and this type of training is usually referred to as internal cross validation in QSAR [81]. A common method is k-fold CV which is used here to explain the model development methodology. K-fold CV divides the training data set into k equally sized subsets and trains the model by using different permutations of these. One subset is always left out of the training and used instead to validate the generated model [82]. This generates k new models with respective performance metrics that have been trained with differing data sets that are used to identify optimal model parameters for the final model trained with the full training data set. The main advantage of this procedure is that over-fitting of the training data is avoided by minimizing the model error. The impact of outliers present in the data is reduced and the final model exhibits better generalization properties [83]. The remaining 20-30% of the samples are used to estimate the final models true accuracy when used on an external data set [83, 84]. There have been many studies where high prediction performance was reported from internal validation but no external validation had been performed and the model later failed to accurately predict new data [85, 86].

Important to remember when dividing the data is that high enough variability of the compounds is present in the data set for training. The data set should contain sufficient distinct excitations in the variables to be able to describe the system behaviour which will be the foundation for the model applicability domain [87]. Therefore, exploratory analysis of the data is recommended as part of the model development, e.g. Principal Component Analysis [88], before dividing the data into training and test sets. This is to ensure that all necessary information will be presented when developing the model as a model cannot extrapolate outside of the boundaries of the data used for its training.

It has been argued that a better approach to the traditional model training regime is by using nested cross validation instead for machine learning methods such as SVM to find the optimal model hyper parameters. Like basic cross validation, nested cross validation has an internal cross validation
loop to train the model. The main difference is that instead of having a static external test set, an outer cross validation loop is used to provide data permutations for training and externally validating the model. This has the advantage of further reducing the model bias by using the full data set more efficiently but at the expense of being more computationally heavy [89].

However, caution needs to be exercised when dealing with small data sets as division into test and training sets can lead to loss of information when developing the model. In such cases, external validation based on the original data set will be difficult where selection of test data points might introduce bias into the model due to data sparsity. Bootstrap resampling which was first introduce by Bradely Efron in 1979 [90], has been shown to be a good method to counter this [91]. In bootstrapping, new data sets are generated by randomly selecting samples from the original data set where sample repetition is allowed. The selection continues until the new data set is of the same size of that of the original. The samples that are not selected in the new data set are used for validation instead. Due to the random selection, many unique permutations are generated which means that a more efficient training of the model with the available data set as well as information about descriptor uncertainty can be generated by investigating the contribution of each descriptor to the model performance of all models [92]. Another problem with small-sample sizes is that due to sparsity the sample population might not give accurate descriptor distribution estimates needed to classify the samples. Preferably, data generated from experimental designs would ensure high data excitation being present in the data set.

Another approach instead of traditionally generating a single model as described above is to use an ensemble or aggregation of generated sub-models for prediction such as bootstrap aggregation (bagging) which was introduced by Leo Breiman [93]. For regression this would generate a mean value and for classification the most voted class would be chosen based on the used sub-models. Bagging also shows the uncertainty of predictions which can be assessed by investigating the prediction variation of all sub-models. This might be preferable if a single model shows high instability in prediction whereas model aggregation potentially can increase the model precision [94]. Bagging has been shown to be particularly efficient in training QSAR models in chromatography settings [40, 41, 95]. It has been argued however that caution needs to be exercised if specific sub-models are selected for the model aggregation which can introduce bias and reduce the models ability to generalize [83].

Model Performance

There are numerous metrics that can be used to evaluate the model performance. In regression, the most commonly used metric is $R^2$ or otherwise known as the goodness of fit of the model which ranges from zero where no output variation is explained, to one where all variation is explained [86]. Another important metric is the mean square error (MSE) which is a measure of the difference between the measured and predicted values. This metric is especially important when accounting for the model complexity in the internal cross validation e.g. selection of model parameters. The MSE for cross validation (MSECV) is generated from the differences between the predicted values and the measured samples in the validation sets. The MSE can be broken down into the three components: the squared model bias, the model variance (or the random error in the fitted model) and an irreducible error which will always be present [96]. The first two terms can be used for the bias-variance trade-off when optimizing the model fit. This means that the model parameters are chosen to minimize the sum of squared model bias and model variance in the MSECV. This ensures the highest generalization for the model with the specific cross validation method that was used.

Metrics used for classification are different when compared to those of regression as the model fit might not necessarily correlate to the number of correctly predicted samples. Instead, metrics are derived from the classification confusion matrix are preferably used. This matrix is a square matrix...
where the dimensions equal the number of assigned classes. The error rate, which is the number of misclassified sample gives a better representation for optimizing the model performance through cross validation compared to that of MSE is used in regression [97]. A popular metric that summarizes the different aspects in the confusion matrix is the Matthew’s correlation coefficient (MCC). It can be used to benchmark different models and it ranges from zero to one, meaning no correlation between descriptors and output to full correlation respectively [98].

Even with descriptor selection the number of descriptors that is used in the model training might still be extensive. In QSAR, Y-randomization has gained a lot of popularity and is used to investigate if random permutations of descriptors contribute to the model performance. The method works by randomizing the output between the samples and a model is generated as described above with the randomized data set. If random descriptor contributions affect the model performance in the external test set, then metrics like the goodness of fit (R²) for regression and MCC for classification will show high correlation (close to one). This means that either further descriptor selection needs to be performed or more samples are needed to increase the available information in the data set. If the values are close to zero however, this means that little to no random contribution from the descriptors occur and that the contributing descriptor from the original model have high correlation to the measured outputs [99].

Selecting a modelling method to be used in QSAR studies can be challenging. Usually trial and error prevail to select the model with highest accuracy and smallest error in order to get a good fit, which is usually done in the form of model benchmarking. To summarize, selection and training of the model must be executed with care as no one method or combination of methods works for all situations. If possible, the impact of the contributing descriptors on the model output should be investigated to see if there is a mechanistic explanation pertained to the descriptors and output, thus adding more credibility to the model. It is therefore important to remain critical of the generated models and thoroughly investigate their stability, precision and if the models make sense [100].

An extensive, but non-exhaustive list of popular modelling methods that has been used in both QSAR and multivariate data analysis in process development is presented in Table S2 where advantages and disadvantages are listed for each method. This review however, is meant to inspire a greater understanding and good practices for developing models and the different aspects to consider when using QSAR for proteins such as mAbs. Therefore, further discussions or comparison of the different modelling methods is out of the scope if this review.

Towards antibody process development by bridging QbD and QSAR

There have been significant advances in computational prediction methods and they are starting to become more common in process development [101]. As mentioned by Zurdo et al. [8], the ability to predict product related characteristics that strongly relate to the QTPP and/or CQAs can greatly simplify process development, especially in the early stages when the product or process knowledge is limited. Implementation of QSAR in process related areas such as protein purification have been researched extensively, e.g. Chen et al. [95], Yang et al. [41], Ladiwala et al. [102], Woo et al. [103] and Robinson et al. [65] to mention a few. Though not all the mentioned examples concern mAbs specifically, the outlined methodology used in the different research articles is still applicable to antibodies. Given the significant proportion of mAb production cost that is incurred during downstream processing, considerable advantages can be gained by being able to predict the performance of chromatographic columns and their effect on product quality early in the process development. In the case of antibodies much of the cost is incurred during the purification due to the strict regulations surrounding clinical safety of the end product [5, 104].
A proposed integration of QSAR into QbD based on these concepts is illustrated in Figure 3 which also shows how the QbD framework can add to and improve the QSAR modelling with addition of new data.

![Figure 3](image)

**Figure 3.** Illustrates the integration of QSAR into QbD where the upper half of illustrates the simplified framework of QbD whereas the lower half illustrates a simplified version of the QSAR framework. The yellow arrows represent transfer of characterisation data from previous mAb processes that can be used directly for model development using QSAR. Developed QSAR models can be used to directly aid in assessing CQAs (red arrows) as well as provide insight into process parameters (grey arrows).

Two main approaches of integrating the QSAR framework into the QbD paradigm can be considered. The first approach is by only using generated structural descriptors for development of models able to predict protein behaviours. An example of this was published by Obrezanova et al. [57] where the authors developed a model being able to predict the probability of mAb aggregation based on the structure of the primary sequence. The method is however more constrained as it requires data generated from identical experimental setups, and therefore identical PP settings in order to assume that the observed effect is caused only by the differences in structure between the proteins. Therefore, models developed this way are better for assessing the manufacturing feasibility and/or potential CQAs before starting the process development (seen as red arrow in Figure 3). The second approach is to use the PPs of interest, taken from previous mAb processes to use directly in the model development (seen as yellow arrows in Figure 3) by either 1) by adding the PPs together with the generated structural descriptors as inputs [105] or 2) structural descriptors are calculated to be dependent on the PPs, meaning that the values of the descriptors will change with changing values of the PPs [41]. The latter is easiest done by generating descriptor from MD simulations where changes in the soluble environment can be implemented. This however requires that data is gathered from similar experimental setups where only the PPs of interest have been varied. This would usually not be a problem when gathering historic data generated from the QbD paradigm as it will often conform to experimental designs based on DoE where the experimental environment is strictly controlled. The added benefit of this approach is that the developed model will be able to account for both the structural differences as well as the impact from the studied PPs when predicting protein behaviour. This can potentially have great value in process
development of new antibodies as PP ranges can be assessed in silico and therefore greatly aid in reducing the number of needed experiments, seen as grey arrows in Figure 3.

The methods described above provide a reference for further risk assessment and characterization to be performed in the QbD framework, as they provide information, such as the behaviour of the product in different scenarios and increase the product understanding. As additional information from new mAb processes becomes available, models can be improved by expanding the datasets used in the model development. This in turn will aid in providing more accurate predictions due to lowering the sparsity by incorporating more protein structures. Available characterization studies from academia can also be used as additional sources of data in order to improve the models by expanding the dataset for model development.

Case study - Aggregation

To illustrate the integration in a more practical way, the QSAR implementation is discussed step by step using a case study to investigate the aggregation of mAbs. Aggregation is a common CQA that is usually investigated as the presence of aggregates are known to cause adverse effects and thus needs to be avoided or removed in process development [106, 107]. Refolding of aggregated mAbs can be performed in some cases by introducing extra processing steps to refold the protein into the desired conformation. However, this will introduce further development and manufacturing costs and prolong the production time. As mentioned before, much of the manufacturing cost is incurred in the purification of the mAb.

The purification is also a stage where the mAb is exposed to more extreme environmental variations. An example of this is the protein A elution step with low pH (2.5-4.0) to elute bound mAb proteins from the chromatographic column. The low pH also allows for virus inactivation which makes this a natural step after elution from protein A [108]. However, it has been shown that low pH promotes faster protein aggregation of mAbs due to conformational changes or modifications of the structure [109, 110]. Thus, this makes an ideal case study for QSAR to investigate the impact of differing protein structures on the aggregation during low pH.

1) Aggregation data can be collected from several mAb processes usually available in past characterization studies from either size exclusion chromatography (SEC) or polyacrylamide gel electrophoresis (PAGE). For the sake of the case study the example is based on data from SEC where the retention profiles and times of high molecular weight species of the mAb aggregates are retained. The tested pH ranges are retained and used as inputs (as well as data of other tested process parameters in real case studies) in the model development in order to investigate their influence on the aggregation.

2) Descriptors are generated according to the flow chart in Figure 2. An initial approach would be to generate descriptors according to the sequence based approach first e.g. with amino acid scale and descriptor generating software. The reason for this is that this approach is much faster compared to that of the homology/MD based approach, but also that adequate models might be generated using the sequence based approach. It should be remembered, however, that descriptors generated from MD simulation might be more powerful in explaining the underlying mechanism of the aggregation as they are generated based on structural properties of the 3D models.

3) Cleaning and pre-treatment is often necessary before the data can be used for modelling. This includes removal of descriptors and process parameters that are static or have missing values. Usually, the different variables will have unique ranges with differing magnitudes. In these cases, pre-treatment in form of autoscaling should be applied to centre the descriptors around
zero and scale them to have a standard deviation of one. This is necessary in order to prevent the offset from the origin and variables with large magnitudes to overpower the model and to give all descriptors equal chance to contribute to the model output [111]. Variable reduction methods can also be applied here to reduce the total number of variables used in the modelling by either using decomposition methods such as PCA or correlation studies. However, models should be generated from both the reduced and full data sets in order to evaluate if the reduction improves the model performance.

4) Benchmarking of different modelling methods, potentially in combination with descriptor selection techniques, should then be carried out to find the optimal modelling solution for the problem statement. As mentioned, it is critical to perform proper evaluation of the model performance in order to select a model that has good prediction on an external test set and can be used for future predictions. A recommendation would be to use Y-randomization to investigate if there is an underlying pattern in the variables that is correlated to the output. The use of bagging can be especially beneficial in this case and can provide valuable information by investigating the uncertainty of the predictions with samples from an external test set or new samples acquired in the future. In bagging the prediction will be the averaged value based on the predictions from all sub-models for a specified structure and process parameter input. From this, further information can be gained by constructing confidence intervals that are computed from the individual predictions of all sub-models and provide a measure of how certain or uncertain the prediction is.

In this way, it is possible to assess the influence of the process parameters on the prediction in more detail. In case of high uncertainty, the distribution of the predictions will span a wider range, which means that further characterization studies are needed to estimate the true aggregation probability. This information can later be added to the training of the model in order to improve it. In cases of low uncertainty, the range of the prediction distribution will be tighter, meaning that all sub-models give similar predictions and characterization studies might not be necessary in these cases as the underlying cause is well understood by the model.

Through this example, we have shown how to integrate the QSAR framework with the QbD paradigm to assess the impact of pH (and potentially other process parameters) and the structure of mAbs on the aggregation propensity. Many other examples from upstream and downstream could also be used to evaluate the impact of structural differences in mAbs on CQAs and CPPs in the process development. For example, media design characterization data such as mentioned in the work of Rouiller et al. [26] could be used to find similarities in media composition related to structural features of different mAbs. Similarly, optimal cell lines for expressing the mAb candidates may be predicted using data as that generated by [25]. Data generated from high-throughput screening of chromatographic columns such as mentioned in the works of Bhambure et al. [24] and Bhambure and Rathore [30] could be used together with QSAR for the prediction of optimal separation and selection of columns. These are a few examples that are possible with emerging technologies that can be applied for risk assessment of CQAs and PPs using QSAR. The structural similarity of the mAbs is of great advantage and it simplifies the analysis of structural differences as compared to other proteins, which makes them ideal for analysis with QSAR.

Conclusions

In an environment where the regulatory standards are becoming increasingly tighter due to increased clinical safety requirements, the biotech industries are forced to raise the bar on quality assurance by introducing the QbD paradigm. QbD exploits process understanding to monitor and control only the CPPs and raw material attributes that affect CQAs of the product and thus mainly focuses on process
understanding. However, a significant limitation of the current implementation of this framework is the inability to quantitatively predict the potential processing risks based on the intrinsic properties of the pharmaceutical product, as in contrast to the (semi-)empirical risk assessment approaches that are used to date. It has been highlighted that considering the intrinsic properties of the product while exploiting all the data generated during the process development of different candidates for risk and process alternatives assessment, inherits potential for significant cost and time savings, while being fully in line with the QbD idea.

A possible approach that can aid to exploit the potential and reduce the current limitations by integrating product knowledge is the QSAR framework. The idea is to link the structural differences within the products to the process characterization data to find patterns between structural differences and product properties that can be used for risk assessment of product behaviour and process operating conditions. With the advancements of in silico methods, their ability to predict complex system properties and the impact on product quality, these methods may be used to effectively reduce the number of experiments needed to characterize the design space. By using the outlined methods to quantify the differences between mAbs and linking those to the process conditions by applying the QSAR methodology, process development could be speeded up and costs be reduced.

The QSAR approach was highlighted as a good candidate to be used together with QbD in order to increase both product and process understanding but also to make the implementation of QbD more effective. The current rate of improvement increases the likelihood that in silico methods will become an integral part in process development.

Acknowledgements

The authors would like to thank Tibor Nagy, Subject Matter Expert, and Graham McCreath, Head of Process Design, from Fujifilm Diosynth Biotechnologies for their valuable discussions. We would also like to thank the Horizon 2020 Marie Skłodowska-Curie actions, grant number 643056, for sponsoring the BioRapid project and making this research possible.

References


**Supplementary Information**

Table S1. List of popular software used in homology modelling and Molecular Dynamics simulations that has been used for antibodies.

<table>
<thead>
<tr>
<th>Homology Modelling Software</th>
<th>Developed by</th>
<th>Accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modeller</td>
<td>University of California, San Francisco</td>
<td>Free</td>
</tr>
<tr>
<td>Molecular Operating Environment (MOE)</td>
<td>Chemical Computing Group</td>
<td>Commercial</td>
</tr>
<tr>
<td>PIGS</td>
<td>University of Rome &quot;Sapienza&quot;</td>
<td>Free</td>
</tr>
<tr>
<td>RosettaAntibody</td>
<td>Gray Lab at John Hopkins University</td>
<td>Free</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecular Dynamics (MD) Simulation Software</th>
<th>Developed by</th>
<th>Accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMBER</td>
<td>University of California, San Francisco</td>
<td>Commercial</td>
</tr>
<tr>
<td>Desmond</td>
<td>Schrödinger</td>
<td>Commercial</td>
</tr>
<tr>
<td>GROMACS</td>
<td>University of Groningen, Royal Institute of Technology and Uppsala University</td>
<td>Free</td>
</tr>
<tr>
<td>MOE</td>
<td>Chemical Computing Group</td>
<td>Commercial</td>
</tr>
<tr>
<td>NAMD</td>
<td>Theoretical and Computational Biophysics Group, University of Illinois</td>
<td>Free</td>
</tr>
<tr>
<td>Protein RECON</td>
<td>Rensselaer Exploratory Center for Cheminformatics Research (RECCR)</td>
<td>Free</td>
</tr>
<tr>
<td>YASARA</td>
<td>YASARA Biosciences</td>
<td>Commercial</td>
</tr>
<tr>
<td>Method</td>
<td>Classification/ Regression</td>
<td>Linear/Non-linear</td>
</tr>
<tr>
<td>-----------------------------------------</td>
<td>----------------------------</td>
<td>-------------------</td>
</tr>
</tbody>
</table>
| Principal Component Analysis (PCA)      | Classification            | Linear            | The PCA method transforms data into orthogonal linear principal components explaining successively reducing levels of variance in the data set. | **Advantages:** The transformation eliminates collinearity and lowers the complexity of the data set by transforming the data into orthogonal principal components. Depending on the data, by plotting the scores of two principal components against each other PCA can be used to identify clusters of data points or gradients.  
**Disadvantages:** The original variables are linear combination of the principal components, wherefore poor performance of PCA on data that contain non-linear combinations can be expected. |
| Linear Discriminant Analysis (LDA)      | Classification            | Both              | LDA tries to find a hyperplane separating the different classes based on a target property. The descriptors are used in a linear combination to define the hyperplane with the help of a so called linear discriminant functions that maximizes the variance between classes and minimizes the variance inside of the class. | **Advantages:** It works well on data that has categorical target properties and continuous descriptor variables. The LDA classifier is also easy to implement.  
**Disadvantages:** Many parameters must be estimated and the model assumes that the data are Gaussian distributed even if this is not the case. The covariance of all classes is also assumed to be the same. |
| k- Nearest Neighbours (k-NN)            | Both                      | Not applicable    | Samples are classified by a distance metric and put into the class that the majority of the samples k nearest neighbours belongs to. | **Advantages:** One of the simplest classifier methods to understand and implement.  
**Disadvantages:** k-NN works best with well separated classes which means that classes that overlap have a probability of being misclassified.  
The integer k needs to be optimized to provide a robust model. The number of samples in each class have to be equal or similar or the model will be biased. |
| Decision trees (DT)                     | Both                      | Non-linear        | DT tries to predict a target class based on several input descriptors. It uses a method called recursive partitioning that divides the data into subsets by adding branching rules based on threshold values for a particular descriptor. | **Advantages:** The trained classifier has a tree-like structure, which gives it very high interpretability as the effects of the descriptors can be seen on the target variable.  
**Disadvantages:** There is usually a problem with over-fitting when using decision trees. This makes pruning necessary to remove branches and make the model more general. Decision trees are also very sensitive to noise and outliers in the data. |
<table>
<thead>
<tr>
<th>Method</th>
<th>Classification/Regression</th>
<th>Linear/Non-linear</th>
<th>Description</th>
<th>Advantages/Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression (MLR)</td>
<td>Regression</td>
<td>Linear</td>
<td>MLR attempts to establish a linear relationship between the input descriptors and the measured output by solving for coefficients for each descriptor.</td>
<td><strong>Advantages</strong>: Due to its simplicity, it is a very easy method to implement and the generated model is highly interpretable. <strong>Disadvantages</strong>: MLR does not work well when there are a lot of collinearity in the descriptor data which can lead to wrongly calculated coefficients. MLR also require more samples than descriptors to work properly.</td>
</tr>
<tr>
<td>Principal Component Regression (PCR)</td>
<td>Regression</td>
<td>Linear</td>
<td>The PCR method tries to overcome the collinearity problem in the descriptor data by use of dimension reduction such as in PCA. The generated independent principal components are then used to establish a linear relationship with the measured output similar to MLR.</td>
<td><strong>Advantages</strong>: In descriptor sets with high collinearity the PCR method can reduce the data set to dependent principal components that can be used on the measured output. <strong>Disadvantages</strong>: The dimension reduction is only dependent on the descriptor input data. Meaning, PCR does not consider the possible relationships between the descriptors and the measured output before the dimension reduction. This can lead to selection of principal components with poor relationship with the measured output.</td>
</tr>
<tr>
<td>Partial Least Squares (PLS)</td>
<td>Regression</td>
<td>Linear</td>
<td>PLS is a method that explains the fundamental relationships between input and output data via a latent variable space of lower dimension, which is constructed such as to maximize the covariance between the inputs and outputs.</td>
<td><strong>Advantages</strong>: It is widely used to construct linear models between input and output data. Like PCA, PLS eliminates collinearity and also has the advantage of being suitable for problems with small sample-sizes. <strong>Disadvantages</strong>: PLS exhibits the same disadvantages as PCA.</td>
</tr>
<tr>
<td>Support Vector Machines (SVM)</td>
<td>Both</td>
<td>Both</td>
<td>When used for regression it is called Support Vector Regression (SVR). SVM simplifies the classification by transforming the original data into a feature space by using kernel functions, which are typically nonlinear. In the feature space the data is easier to separate into different classes.</td>
<td><strong>Advantages</strong>: Two of the main advantages of SVM is 1) its ability to make good predictions even with small sample sizes; and 2) the generalization ability of the model is high. <strong>Disadvantages</strong>: SVM does not have very good interpretability of how descriptors effect the classification due to the transformation in to feature space. SVM also requires a lot of tweaking as to which kernel to use and optimizing parameters.</td>
</tr>
<tr>
<td>Artificial Neural Networks (ANN)</td>
<td>Both</td>
<td>Both</td>
<td>Usually ANN consists of neurons in an input layer, hidden layer and an output layer. The connections between neurons carries a weight that is determined during training, thereby weakening or strengthening the connection as it learns how to connect the input and output data.</td>
<td><strong>Advantages</strong>: ANN is a biologically inspired method and has been successfully applied to areas such as toxicology, pharmacology and physiochemical properties prediction, besides others. The method is able to detect complex relationships between descriptors and activity with the use of multiple layers. <strong>Disadvantages</strong>: There can be problems with over-fitting of the training data set which decreases the models ability to generalize, thus not accurately predicting for samples outside of the training data set. The method can also be computationally expensive if the problem is complex.</td>
</tr>
<tr>
<td>Method</td>
<td>Classification/Regression</td>
<td>Linear/Non-linear</td>
<td>Description</td>
<td>Advantages/Disadvantages</td>
</tr>
<tr>
<td>--------------------</td>
<td>---------------------------</td>
<td>-------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
</tbody>
</table>
| Random Forest (RF) | Both                      | Non-linear        | RF is a technique that uses many decision trees explained above. However, the trees are only trained with subsets of the descriptors and samples, which makes all trees different. Classification is then done by a majority rule over all trees. | **Advantages:** Avoids over-fitting by randomly creating subsets of descriptors and samples. It makes the model more generalized and is considered to be relatively robust.  
**Disadvantages:** Due to the majority rule and that all trees are different the interpretability of the model is considerably lower. RF can also be computational taxing if many trees need to be generated and trained from a big data set. |