Characterization of absorption enhancers for orally administered therapeutic peptides in tablet formulations - Applying statistical learning

To develop a successful oral formulation of insulin for treatment of type-2 diabetes patients would be a great milestone in terms of convenience. Besides protecting insulin from enzymatic cleavage in the small intestine, the formulation must overcome the intestinal epithelia barrier. Absorption enhancers are needed to ensure even a few percent of insulin are taken up. In thesis article 1, various methods to measure the effect of absorption enhancement and enzyme stability of insulin were applied. The major class of absorption enhancers is surfactant-like enhancers and is thought to promote absorption by mildly perturbing the epithelial membranes of the small intestine. The Caco-2 (Carcinoma Colon) cells can grow an artificial epithelial layer, and are used to test the potency of new absorption enhancers. This project was aimed to identify new absorption enhancers, that are both potent and sufficiently soluble. Quantitative structural activity relationship (QSAR) modeling is an empiric approach to learn relationships between molecular formulas and the biochemical properties using statistical models. A public data set testing the potency of absorption enhancers in Caco-2 was used to build a QSAR model to screen for new potent permeation enhancers. Thesis article 2 contains likely the first QSAR model to predict absorption enhancement. The model was verified by predicting molecules not tested before in Caco-2. The Caco-2 model overestimates the clinical effect of lipophilic permeation enhancers. In the Caco-2 model all reagents are pre-dissolved, and therefore the assay cannot predict critical solubility issues and bile salt interactions in the final tablet formulation. A QSAR solubility model was built to foresee and avoid slow tablet dissolution. Due to enzyme kinetics, slow tablet dissolution will allow most insulin to be deactivated by intestinal enzymes. The combined predictions of potency and solubility, will likely provide a more useful in-silico screening of potential permeation enhancers.

Random forest was used to learn relationships between molecular descriptors and potency or solubility. However, unlike multiple linear regression, the explicitly stated random forest model is complex, and therefore difficult to interpret and communicate. Any supervised regression model can be understood as a high dimensional surface connecting any possible combination of molecular properties with a given prediction. This high dimensional surface is also difficult to comprehend, but for random forests, it was discovered that a method, feature contributions, was especially useful to decompose and visualize model structures. The visualization technique was named forest floor and could replace the otherwise widely use technique partial dependence plots, especially in terms of discovering interactions in the model structure. Thesis article 3 describes the forest floor method. An R package forestFloor was developed to compute feature contributions and visualize these according to the ideas of thesis article 3. Better interpretation of random forest models is an exciting interdisciplinary field, as it allows investigators of many backgrounds to find fairly complicated relationships in data sets without in advance specifying what parameters to estimate. Forest floor was used to explain how potency and solubility were predicted by random forest models.