Bakken.[30,31] Balakin et al. modeled a large set of 65,500 molecules with measured DMSO solubility. Molecules were classified as insoluble if they were not soluble at 0.01 mol/l. A Kohonen neural network was able to correctly classify 93% of compounds using only eight descriptors. Such models work by mapping the input data into a smaller dimensional space based on the nodes and making predictions based on node membership. In essence, a molecule is predicted as soluble or insoluble in DMSO based on the neighboring molecules in its assigned node. Surprisingly, a standard neural network performed worse on the same data, having approximately 75% accuracy. At Pfizer, 33,329 compounds dissolved in 30 mM DMSO stock solutions were visually inspected for precipitates. They computed 200 2D descriptors (78 E-state keys and a set of 122 from the MOE software package) to build five models to classify compounds that showed precipitation versus those that showed no precipitation. Test set accuracy was reasonably good across all five models: recursive partitioning 81%, random forest 81%, binary quantitative structure/activity relationship (QSAR) 74%, self-organizing map 69%, and linear discriminant analysis 76%.

Little work has been performed to model solubility while taking into account crystal packing. Johnson et al.[32] published an initial attempt using calculated intrinsic solubility corrected for effects of ionization, and crystal-packing forces derived from an escalating temperature molecular dynamics simulation. Although the model requires crystal structure information, it can be applied to analogs that do not have crystal structures simply by overlaying those analogs onto the known crystal form to begin the simulation. Results suggest this type of model could be useful to understand the solubility of late-stage optimization and early development candidates, although it is highly dependent on pK_a estimates.

INTESTINAL ABSORPTION

Theory and computational aspects of intestinal permeability have been reviewed in detail by Egan and Lauri.[33] A drug must be somewhat permeable through the membrane of the intestinal tract if it is to be administered orally and achieve systemic exposure. The rate of membrane permeability is strongly related to the lipophilicity and hydrophilicity of the molecule.

Egan et al.[33,34] demonstrated that a statistically based classification model built using only PSA and Alog*P*98 could predict the region of chemical space occupied by well-absorbed (>90% absorbed) molecules and exclude poorly absorbed molecules (<30% absorbed). Molecules with absorption in the range 30–90% were not used because of large data variability. Actively transported molecules were excluded. These results were validated on Caco-2 permeability assay data from drug discovery projects at Pharmacopeia. The Caco-2 permeabilities were shown to have a hill-shape in PSA-Alog*P*98 space. The sides of the hill declined rapidly at the edge of the well-absorbed region and less than 10% of highly permeable molecules were outside the well-absorbed region, while only 21% of poorly permeable molecules were inside the well-absorbed region.

In an excellent article, Zhao et al.[35] assembled a carefully reviewed literature set of human absorption data on 241 drugs. They showed that a linear regression model built with five Abraham descriptors could fit percent human absorption data reasonably well ($r^2 = 0.83$, rmse = 14%). The descriptors are excess molar refraction (E), polarizability (S), hydrogen bond acidity (A), hydrogen bond basicity (B), and McGowan volume (V), all related to lipophilicity, hydrophilicity, and size. In a follow-up article, data on rat absorption for 151 drugs was collected from the literature and modeled using the Abraham descriptors.[36] A model with only descriptors A and B had $r^2 = 0.66$, rmse = 15%.

All in vivo data, including the human and rat absorption data used by both Egan and Zhao et al., have considerable variability. Zhao et al. comment that measurements of percent absorbed for the same molecule may vary by 30% and that the 95% confidence interval for a prediction is approximately 30% given a model rmse of 15%. This is approximately the same as the normal experimental error for absorption values. This means that models predicting percent absorbed have to be carefully interpreted (i.e., a prediction of 30% absorbed really means the molecule is predicted to have absorption from 15–45%). For this problem, regression models are really no better than classification models because of the variability in absorption data.

A classification regression tree model using 28 descriptors to predict the fraction absorbed for a large set of 1,260 drugs and drug candidates has been published.[37] The training set was 899 molecules and fraction absorbed was split into six classes (0–0.19, 0.2–0.31, 0.32–0.43, 0.44–0.59, 0.6–0.75, 0.76–1). Predicted values were reported as the median of each class. Average absolute error (AAE) for the test set of 362 molecules was 0.169 and 80.4% of molecules were predicted within one class of their actual class. For 37 proprietary molecules having human data, AAE = 0.14 and 86.4% of molecules were predicted correctly within one class.

Descriptors such as PSA, Clog*P*, and the Abraham descriptors can be interpreted in terms of chemical structure without much difficulty. Jones et al.[38] showed that quantum mechanical descriptors can be used to successfully predict intestinal absorption and at the same time provide an interpretable model. They used the data set of Zhao et al.[35] and computed molecular surface charges using density functional theory. The model quality was almost identical to the Abraham descriptor model reported by Zhao et al. (rmse = 15% for the same test set). The surface charges were mapped to the 3D structure of drugs creating an easily interpretable image.

Intramolecular hydrogen bonds can have an effect on membrane permeability. If a polar molecule can adopt a conformation that forms intramolecular hydrogen bonds, it will be able to present a more lipophilic surface to the