thermodynamic solubility using shake-flask with high-pressure liquid chromatography with UV detector (HPLC-UV) or liquid chromatograph/mass spectrometry detection. These factors can cause a single molecule to have widely differing solubility values that are not comparable.

From a modeling standpoint, the prediction of a molecule's solubility is a very difficult task because of the issues listed above.[16–18] The problem of predicting solubility has been attacked with some success with complex neural network models. Although not interpretable, neural networks can function as a black-box in silico assay. Other techniques that are more interpretable have also been applied to the problem.

An interesting approach for estimating the effects of small modifications on molecular properties such as solubility was published by Leach et al.[19] The technique is called "matched molecular pairs analysis." First, a set of specific structural transformations are used to search a set of molecules having some type of property data. Subsets of almost identical molecules having each transformation are identified (e.g., all molecules differing by *p*-fluorine on a phenyl ring). The percentage of molecules with a positive property value change is computed, and the binomial distribution is used as a statistical test to ascertain if the change is significant. For example, the authors reported that when an amide is methylated, 112 of 142 pairs had increased solubility by an average of +0.64 log units. The percentage of pairs with increased solubility was 79% with a 95% confidence interval of 71–85%, indicating the effect is statistically significant. This technique is not limited to solubility but can be applied to any property of a molecule, ADME or otherwise. The authors also show examples of insights gained from matched molecular pairs analysis of data on protein binding and oral exposure in rats. Matched molecular pairs analysis is clearly interpretable and as the authors state, "can be used as a tool to test many of the 'rules of thumb' that abound within medicinal chemistry."

Another simple approach to classifying molecules as soluble or insoluble was published by Lamanna et al.[20] They used recursive partitioning to classify 3,563 molecules as soluble/insoluble using a small set of descriptors. Multiple models were found which were predictive. The best model used only two simple descriptors: MW and the descriptor "aromatic proportion" and had an accuracy of 81% for a test set of 1,200 molecules using a cutoff of 30 μM.

Huuskonen[21] assembled aqueous solubility data for 1,297 organic molecules and modeled it using neural network and linear regression models trained on 55 connectivity, shape, and electrotopological state descriptors. Test set results were $r^2 = 0.92$ and standard deviation (s) $= 0.60$ for the neural network and $r^2 = 0.88$ and $s = 0.71$ for the linear regression model. Yan et al.[22] were able to build neural network and linear regression models of comparable quality for the Huuskonen data set using only 18 topological descriptors. Test set results were $r^2 = 0.94$ and $s = 0.52$ for the neural network model and $r^2 = 0.89$ and $s = 0.68$ for the linear regression model. Further work by Yan et al.[23] modeled the aqueous solubility of a set of 2,743 drug discovery molecules from Merck KGaA, resulting in a neural network model using 18 2D topological descriptors with $r = 0.92$ and $s = 0.62$. The authors note that the Huuskonen set is limited in diversity in comparison to the Merck KGaA data set.

One problem highlighted by several reviewers[17,24] is that data sets like the Huuskonen set cover unnecessarily large ranges of solubility. The Huuskonen set covers the range log*S* (log of solubility in mol/l) from −11.62 to +1.58, which converts approximately to 9.6×10^{-7} to 1.5×10^{7} μg/ml for a MW of 400 Da. Johnson and Zheng[17] recommend a pharmaceutically relevant range of 0.1 to 250 μg/ml as more appropriate.

However, the issue is more complex than a simple range. Lipinski[25] provides better guidance for minimum acceptable solubility based on maximal absorbable dose calculations. These take into account dose amount and permeability both of which have significant effects on required solubility. For example, the minimum acceptable solubility for a 0.1 mg/kg human dose (a 7 mg pill) of a high-permeability molecule is 1 μg/ml, whereas the minimum acceptable solubility for a 10 mg/kg human dose (a 700 mg pill) of a low permeability molecule is 2,100 μg/ml. This range is somewhat similar to the range recommended by Johnson and Zheng, but it is important for both medicinal chemists and modelers to be aware of the factors modifying the minimum acceptable solubility values within the solubility range relevant for drug discovery.

Goeller et al.[26] at Bayer modeled buffer solubility at pH 6.5 using a data set containing 5,000 molecules whose solubility was measured in a consistent fashion. The Bayer assay was a high-throughput assay starting from DMSO stock diluted to 1% DMSO in phosphate-buffered saline at pH 6.5 and using HPLC detection. The log*S* range is approximately −6 to −3. The model used 65 VAMP/PROPGEN descriptors computed from 3D structures plus eight common 2D descriptors. These descriptors were used to train various neural networks. The best neural network had a root-mean-squared error (rmse) of 0.73 and 83% of predictions had <1.0 log unit error on a test data set of 7,222 molecules.

Recently, Gaussian process nonlinear regression was used to model a set of combined literature aqueous solubility data and shake flask buffer solubility data for 632 molecules at pH 7.0–7.4 from Schering AG.[27] This machine learning algorithm is just beginning to be used in drug discovery modeling. Gaussian process models can provide error estimates for predictions and can automatically select features. Other studies on modeling solubility using Gaussian processes have also been published. The error bars shown in these articles are wide enough to be alarming.[28,29]

As mentioned, solubility in DMSO is important for compound storage and high-throughput screening efforts. Computational models for the prediction of DMSO solubility have been reported by Balakin et al. and Lu and