ANN is capable of self-learning from the training data to maximize the prediction capability. This is particularly useful in applications where the complexity of the data or task makes the design of such a function by hand impractical (Haykin, 1999).

Goller et al. (2006) present an ANN model for the prediction of solubility of organic compounds in buffer at pH 6.5 to mimic the medium in the human gastrointestinal tract. The model was derived from consistently performed solubility measurements of about 5000 compounds. Semiempirical VAMP/AM1 quantum-chemical wave function-derived, HQSAR-derived log *P*, and topology-based descriptors were employed after preselection of significant contributors by statistical and data mining approaches. Ten ANNs were trained, each with 90% as a training set and 10% as a test set, and deterministic analysis of prediction quality was used in an iterative manner to optimize ANN architecture and descriptor space on the basis of Corina three-dimensional molecular structure and AM1/COSMO single point wave function. In production mode, a mean prediction value of the 10 ANNs was created, as was a standard deviation-based quality parameter. The productive ANN based on Corina geometries and AM1/COSMO wave function resulted in an r^2 cv of 0.50 and a root-mean-square error of 0.71 log units, with 87% and 96% of the compounds having an error of less than 1 and 1.5 log units, respectively. The model was able to predict permanently charged species, for example, zwitterions or quaternary amines, and problematic structures such as tautomers and unresolved diastereomers as well as neutral compounds.

Tantishaiyakul (2005) developed a model to predict aqueous solubility of benzylamine salts using multivariate PLS and ANN. Molecular descriptors, including binding energy and surface area of salts, were calculated by the use of Hyperchem and ChemPlus QSAR programs for Windows. Other physicochemical properties, such as hydrogen acceptors for oxygen and nitrogen atoms, hydrogen-bond donors, hydrogen-bond forming ability, molecular weight, and calculated log partition coefficient (clog *P*) of *p*-substituted benzoic acids, were also used as descriptors. In this study, the predictive ability of ANN, especially multilayer perceptron (MLP) architecture networks, was found to be superior to PLS models. The best ANN model derived, a 6-1-1 architecture, had an overall R^2 of 0.850 and root-mean-square error for cross-verification and test set of 0.189 and 0.185 log units, respectively. Since all the utilized descriptors are readily obtained from calculation, these models offer the advantage of not requiring experimental determination of some descriptors.

Jouyban et al. (2004) applied ANN to calculate the solubility of drugs in water–cosolvent mixtures, using 35 experimental datasets. The networks employed were feedforward back-propagation errors with one hidden layer. The topology of neural network was optimized in a 6-5-1 architecture. All data points in each set were used to train the ANN and the solubilities were back-calculated employing the trained networks. The difference between calculated solubilities and experimental values was used as an accuracy criterion and defined as mean percentage deviation (MPD). The overall MPD and its SD obtained for 35 datasets were 0.90% ± 0.65%. To assess the prediction capability of the method, five data points in each set were used as a training set and the solubility at other solvent compositions was predicted using trained ANNs, whereby the overall MPD (±SD) for this analysis was 9.04% ± 3.84%. When all the 496 data points from 35 datasets were used to train a general ANN model, the MPD (±SD) was 24.76% ± 14.76%. To test the prediction capability of the general ANN model, all data points with odd set numbers from 35 datasets were employed to train the ANN model, and the even numbered data were predicted with an overall MPD (±SD) of 55.97% ± 57.88%. When an ANN model was developed for a given cosolvent system, the overall MPD was smaller than 10%. When the ANN results were compared with those obtained from the most accurate multiple linear regression model, namely, the combined nearly ideal binary solvent/Redlich–Kister equation, it showed that ANN was superior to the regression model.

Yan et al. (2004) developed several quantitative models for the prediction of aqueous solubility of organic compounds, on the basis of a diverse dataset with 2084 compounds, by using multilinear regression analysis and back-propagation neural networks. The compounds were described by two different structure representation methods: (1) with 18 topological descriptors and (2) with 32 radial distribution function codes representing the three-dimensional structure of a molecule