The Advent of Generative Chemistry
Quentin Vanhaelen, Yen-Chu Lin and Alex Zhavoronkov

Legend Table 1: A summary of common terms in machine learning

-Machine learning (ML): ML refers to algorithms that learn from and make predictions on data by building a model from sample inputs. ML employed for computing tasks where designing and programming explicit algorithms with good performance is difficult or infeasible. Today, most common traditional ML methods are k-nearest neighbors (kNN), logistic regression (LR), support vector machines (SVM), gradient boosting machines (GBM), and random forest (RF). The performance of ML methods can vary depending on the type of task (regression or classification), types, and amount of data to handle.
-Deep learning (DL): DL refers to a class of ML techniques that exploit many layers of non-linear computational units to model complex relationships among data. These architectures, composed of multiple layers, are commonly called deep neural networks (DNNs), or sometimes stacked neural networks. The difference between the single-hidden-layer artificial neural networks (ANNs) and DNNs is the depth; that is, the number of layers of nodes through which data is processed. Usually, more than three layers (including input and output) qualify as "deep" learning. Thus, "deep" is a technical term that means more than one hidden layer. DNNs use a cascade of many layers of nonlinear processing units for feature extraction. Each successive layer uses the output from the previous layer as input. Higher level features are derived from lower level features to form a hierarchical representation. This hierarchy of features is called a deep architecture. These methods are capable of learning multiple levels of representations that correspond to different levels of abstraction. These levels form a hierarchy of concepts.
-Autoencoder (AE): An AE contains an encoder part, which is a neural network to transform the information received from the input layer to the hidden units, and then couples a decoder neural network with the output layer having the same number of nodes as the input layer. The purpose of the decoder neural network is to reconstruct its own inputs from a fewer number of hidden units and thus AE are used for nonlinear dimensionality reduction.
-Generative Adversarial Networks (GANs): GANs are structured, probabilistic models for generating data. Being an unsupervised technique, GANs can be used to generate data similar to the dataset that the GAN was trained on. A GAN consists of two DNNs called Discriminator and Generator. The discriminator estimates the probability that a given sample is coming from the real dataset. It works as a critic and is optimized to distinguish the fake samples from the real ones. The generator outputs synthetic samples using a noise variable as input following a distribution. It is trained to capture the real data distribution so that it can generate samples with distribution which are as real as possible. The generator should improve its output until the discriminator is unable to distinguish the generated output from the real ones. The two models compete against each other during the training process. The goal of the generator is to try to trick the discriminator, while the discriminator attempts to not be cheated. This process happening between the two models motivates them to improve their functionalities in order to obtain generated samples indistinguishable from the real data.
-Reinforcement learning (RL): RL refers to goal-oriented algorithms, which learn how to attain a complex objective or maximize along a particular dimension over many steps. RL algorithms operate in a delayed return environment, where it is not straightforward to figure out which action leads to which outcome over many time steps. Thus, RL aims at correlating immediate actions with the delayed returns they produce. The reinforcement takes place in the sense that RL algorithms are penalized when making the wrong decisions, and they get rewarded when making the right one. RL algorithms are expected to increase performance in more ambiguous, real-life environments.