Explainable artificial intelligence: evolution, achievements and perspectives

Pavel POLISHCHUK

Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University, Hnevotinska 5, 77900, Olomouc, Czech Republic

From the very beginning QSAR models followed human intuition. They used clearly interpretable descriptors and simple methods to establish structure-activity relationships. This made these models transparent and easy to understand for researchers. Information retrieved from such models could help to understand the underlying mechanisms or could be used to guide further design steps of compounds with improved properties. Further development of QSAR models was focused on improvement of their predictive ability that was successfully achieved by introduction of novel descriptors and machine learning methods. That made the resulting models more obscure due to their very complex internal structure or using non-interpretable descriptors. These models started to be considered as “black boxes” and a popular belief appeared that a trade-off between predictivity and explainability of models exists. This was true to some extent until the development of interpretation approaches which could estimate contributions of atoms or fragments directly from models. In theory, these approaches are applicable to any kind of predictive models that makes all such models interpretable.

More recently, neural networks gained much attention and wide applicability in many fields including chemoinformatic. These models are used to predict properties of molecules, reaction outcomes and optimal conditions, generate new chemical entities with desired properties, etc. Due to the high complexity of neural networks they could capture hidden biases in datasets or spurious correlations that lower the credibility in them. Therefore, the interest in interpretation of neural networks arose greatly last years. Multiple approaches were suggested and many of them were adapted in chemoinformatics. Some of them are model-specific, others have a wider applicability. The large number of these approaches makes it difficult to choose a proper method for interpretation of a particular model. Therefore, certain steps were performed to create specific data sets to benchmark existing and developing approaches. It was demonstrated that not all interpretation approaches were able to retrieve proper structure-activity relationships and their careful investigation is required. However, approaches which were developed previously are also applicable to explain decisions of neural network models.

Despite of all successes in explainability of complex machine learning models the interpretation is still in infancy and not widely used in research work. However, it may provide great advantages and feedback for specialist beyond the chemoinformatic field. Therefore, integration of these approaches in research pipelines and their active use for making decisions can be considered as a major challenge for explainable artificial intelligence in the near future.

This work was funded by the INTER-EXCELLENCE LTARF18013 project (MEYS), the European Regional Development Fund - Project ENOCH (No. CZ.02.1.01/0.0/0.0/16_019/0000868), ELIXIR CZ research infrastructure project (MEYS Grant No: LM2018131).