ABSTRACT

Iovanac, Nicolae C. PhD, Purdue University, July 2021. Generative, Predictive, and Reactive Models for Data Scarce Problems in Chemical Engineering. Major Professor: Brett M. Savoie

Data scarcity is intrinsic to many problems in chemical engineering due to physical constraints or cost. This challenge is acute in chemical and materials design applications, where a lack of data is the norm when trying to develop something new for an emerging application. Addressing novel chemical design under these scarcity constraints takes one of two routes: the traditional forward approach, where properties are predicted based on chemical structure, and the recent inverse approach, where structures are predicted based on required properties. Statistical methods such as machine learning (ML) could greatly accelerate chemical design under both frameworks; however, in contrast to the modeling of continuous data types, molecular prediction has many unique obstacles (e.g., spatial and causal relationships, featurization difficulties) that require further ML methods development. Despite these challenges, this work demonstrates how transfer learning and active learning strategies can be used to create successful chemical ML models in data scarce situations.

Transfer learning is a domain of machine learning under which information learned in solving one task is *transferred* to help in another, more difficult task. Consider the case of a forward design problem involving the search for a molecule with a particular property target with limited existing data, a situation not typically amenable to ML. In these situations, there are often correlated properties that are computationally accessible. As all chemical properties are fundamentally tied to the underlying chemical topology, and because related properties arise due to related moieties, the information contained in the correlated property can be leveraged during model training to help improve the prediction of the data scarce property. Transfer learning is thus a favorable strategy for facilitating high throughput characterization of low-data design spaces.

Generative chemical models invert the structure-function paradigm, and instead directly suggest new chemical structures that should display the desired application properties. This inversion process is fraught with difficulties but can be improved by training these models with strategically selected chemical information. Structural information contained within this chemical property data is thus transferred to support the generation of new, feasible compounds. Moreover,

this transfer learning approach helps ensure that the proposed structures exhibit the specified property targets. Recent extensions also utilize thermodynamic reaction data to help promote the synthesizability of suggested compounds. These transfer learning strategies are well-suited for explorative scenarios where the property values being sought are well outside the range of available training data.

There are situations where property data is so limited that obtaining additional training data is unavoidable. By improving both the predictive and generative qualities of chemical ML models, a fully closed-loop computational search can be conducted using active learning. New molecules in underrepresented property spaces may be iteratively generated by the network, characterized by the network, and used for retraining the network. This allows the model to gradually learn the unknown chemistries required to explore the target regions of chemical space by *actively* suggesting the new training data it needs. By utilizing active learning, the create-test-refine pathway can be addressed purely *in silico*. This approach is particularly suitable for multi-target chemical design, where the high dimensionality of the desired property targets exacerbates data scarcity concerns.

The techniques presented herein can be used to improve both predictive and generative performance of chemical ML models. Transfer learning is demonstrated as a powerful technique for improving the predictive performance of chemical models in situations where a correlated property can be leveraged alongside scarce experimental or computational properties. Inverse design may also be facilitated through the use of transfer learning, where property values can be connected with stable structural features to generate new compounds with targeted properties beyond those observed in the training data. Thus, when the necessary chemical structures are not known, generative networks can directly propose them based on function-structure relationships learned from domain data, and this domain data can even be generated and characterized by the model itself for closed-loop chemical searches in an active learning framework. With recent extensions, these models are compelling techniques for looking at chemical reactions and other data types beyond the individual molecule. Furthermore, the approaches are not limited by choice of model architecture or chemical representation and are expected to be helpful in a variety of data scarce chemical applications.