Report III : Molecular machine learning with conformer ensembles

Irem Begüm Gündüz - 7026821

Contents

Overview of the Approach

A potential in silico regularly utilized in the drug development process is virtual screening. Scientists can use the program to find potential candidates for experimental investigation. Thus, the cost and time required to bring a novel drug to market can be decreased by automatically identifying promising new drug candidates. In the human body, molecules often adopt a particular three-dimensional configuration referred to as conformation. These molecules are not static. In other words, these molecules are able to rotate or spin since they are not in rigid form. This paper claims that a molecular ensemble is definable by compiling all the permitted conformational mobility. This suggestion is significantly important because it enables us to examine a more accurate model of the molecule in its environment.

Current approaches are used to establish three-dimensional conformations of molecules, separated into two categories: physics-based simulations and, data-driven supervised models. As physics-based simulations, computational docking and molecular simulation are constrained by scoring functions and force fields, whereas the performance of the machine learning methods' performance depends on the availability of data. Methods of supervised learning may be preferred because of the advantages listed below: In theory, a neural network can learn any function. Supervised prediction is a significant upgrade over physics-based techniques in terms of speed and performance. For example, neural networks can learn how to map molecules based on their physical properties.

QSAR or classical QSAR (quantitative structure-activity relationship) is a commonly used traditional methodology to predict molecular properties. The generation of features by traditional QSAR approaches is based on physical constraints. One such hypothesis is that the interactions between ligands and receptors impact binding affinity and specificity. For instance, the well-known *CoMFA* approach aligns molecules in a 3D grid, calculates the steric and electrostatic fields at various locations surrounding the molecules, aggregates these values into features, and then uses partial least-squares to link the features with biological activity. Cramer, Patterson, and Bunce (1988) In addition to these, the *GRID* approach takes hydrogen bonding and hydrophobic interactions into consideration. Goodford (1985) These are the commonly used 3-D QSAR methodologies. There are additional 4D QSAR techniques that utilize various molecular conformations.

In addition to QSAR, deep learning is one of the broad approaches used in computational drug discovery. Due to rapid advances in representation learning, such as graph convolutional neural networks, deep learning has been gaining attention over the years. Thus, the use of deep learning methodologies for molecular property prediction is also increased. Both QSAR and deep learning approaches forecast molecular properties by applying a readout function to a molecule's characteristics. In this context, message-passing neural networks gained additional popularity as they get their features directly from the molecules in the training set.

"Can three dimensional information from numerous conformers enhance molecular properties that predicted by deep learning methods?" is the main research question that this work seeks to answer. Multiple deep learning models have been trained to address this question by incorporating multiple conformer inputs and

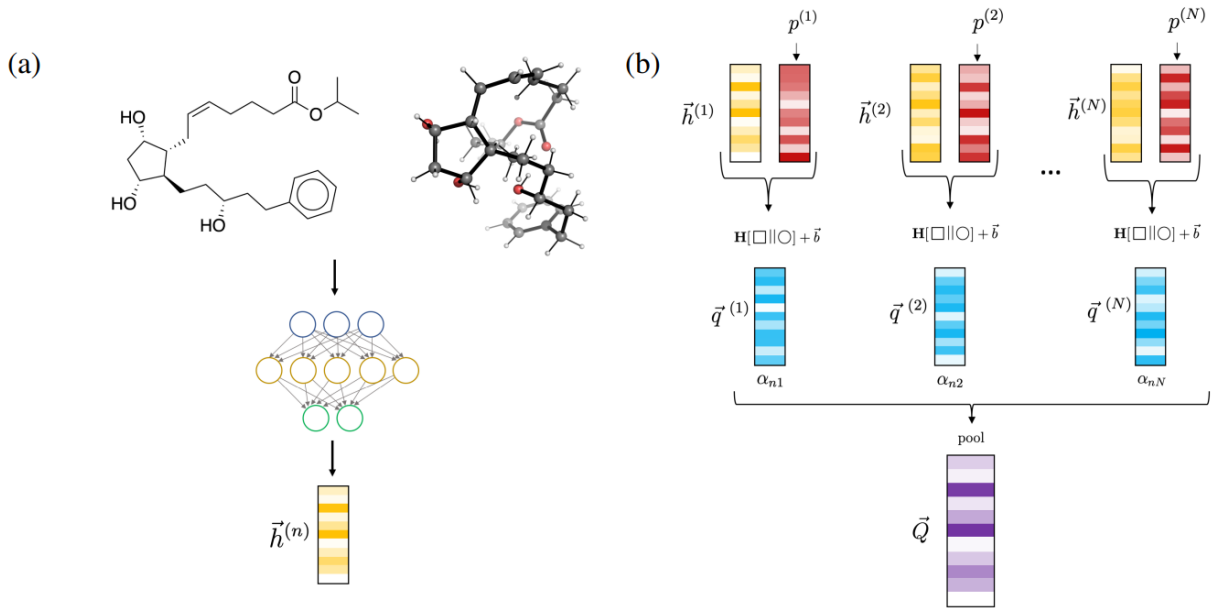

Figure 1: Overwiew of the methdology.

conformer attention to key architectures such as *ChemProp* and *Schnet*. Axelrod and Gomez-Bombarelli (2020)

Discussion

References

Axelrod, Simon, and Rafael Gomez-Bombarelli. 2020. "Molecular machine learning with conformer ensembles," December. https://doi.org/10.48550/arxiv.2012.08452.

Cramer, Richard D., David E. Patterson, and Jeffrey D. Bunce. 1988. "Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins." *Journal of the American Chemical Society* 110 (18): 5959–67. https://doi.org/10.1021/ja00226a005.

Goodford, P. J. 1985. "A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules." *Journal of Medicinal Chemistry* 28 (7): 849–57. https://doi.org/10.1021/jm00145a002.