De novo PROTAC design using graph-based deep generative models

Submitting to: AI4Science Workshop 2022

PROteolysis TArgeting Chimeras (PROTACs) are an emerging therapeutic modality that degrade a protein of interest (POI) by marking it for degradation by the proteasome. They often take on a three-component barbell-like structure consisting of two binding domains and a linker. While a promising modality, it can be challenging to predict whether a new PROTAC will lead to protein degradation as that is dependent on the cooperation of all subunits to form a successful ternary structure. As such, PROTACs remain a laborious and unpredictable modality to design because the functionalities of each component are highly interdependent. Recent developments in artificial intelligence (AI) suggest that deep generative models can assist with the *de novo* design of molecules displaying desired properties, yet their application to PROTAC design remains largely unexplored. Additionally, while previous AI-based approaches have optimized the linker component given two active domains, generative models have not yet been applied to optimization of the other two – the warhead and E3 ligand.

Here, we show that a graph-based deep generative model (DGM) can be used to propose novel PROTAC structures. The DGM follows the approach of GraphINVENT, a gated-graph neural network which iteratively samples an action space and formulates a sequence of steps to build a new molecular graph. We also demonstrate that this model can be guided towards the generation of PROTACs that are predicted to effectively degrade a POI through policy gradient reinforcement learning (RL). Rewards during RL are applied based on a boosted tree surrogate model that predicts a PROTAC's degradation potential for a specific POI, showing that a nonlinear scoring function can fine-tune a deep molecular generative model towards desired properties. Using this approach, we achieve a model where activity against IRAK3 (a pseudokinase implicated in oncologic signaling) is predicted for >80% of sampled PROTACs after RL, compared to 50% predicted activity before any fine-tuning.