Aplicacion de herramientas bioinformaticas e Inteligencia Artificial para el desarrollo de nuevos Farmacos

Adolfo Centeno Tellez
Departmento de Ciencias Computacionales
Instituto Tecnologico y de Estudios Superiores de Monterrey
Campus Puebla, Mexico

Gonzalo Aranda Abreu
Centro de Investigacion en Ciencias Cerebrales
Universidad Veracruzana
Campus Xalapa, Mexico

March 6, 2021

Abstract

Generative models are becoming a tool of choice for exploring the molecular space. These models learn on a large training dataset and produce novel molecular structures with similar properties. Generated structures can be utilized for virtual screening or training semi-supervised predictive models in the downstream tasks. While there are plenty of generative models, it is unclear how to compare and rank them Este trabajo propone la creacion de un marco de trabajo basado en Inteligencia Articial para el descubrimiento de nuevos medicamentos, usando el metodo insilico.

1 Introduction

En la actualidad el uso de las grandes clusters de supercomputadoras, la inteligencia artificial particularmente el Deep Learning han ayudado en la

generacion de conocimiento nuevo en diversas areas de la ciencia. El presente trabajo pretende usar redes neuronales artificiales del tipo Variational Autoencoders para el descubrimiento de nuevos medicamentos usando la base de datos MOSES como principal fuente de informacion para el entrenamiento.

In this work, we introduce a benchmarking platform called Molecular Sets (MOSES) to standardize training and comparison of molecular generative models. MOSES provides a training and testing datasets, and a set of metrics to evaluate the quality and diversity of generated structures. We have implemented and compared several molecular generation models and suggest to use our results as reference points for further advancements in generative chemistry research. The platform and source code are available at https://github.com/molecularsets/moses.

The discovery of new molecules for drugs and materials can bring enormous societal and technological progress, potentially curing rare diseases and providing a pathway for personalized precision medicine [1]. However, complete exploration of the huge space of potential chemicals is computationally intractable; it has been estimated that the number of pharmacologically-sensible molecules is in the order of 1023 to 1080 compounds [2, 3]. Often, this search is constrained based on already discovered structures and desired qualities such as solubility or toxicity. There have been many approaches to exploring the chemical space in silico and in vitro, including high throughput screening, combinatorial libraries, and evolutionary algorithms [4–7]. Recent works demonstrated that machine learning methods can produce new small molecules [8–11] and peptides [12] showing biological activity.

Over the last few years, advances in machine learning, and especially in deep learning, have driven the design of new computational systems for modeling increasingly complex phenomena. One approach that has been proven fruitful for modeling molecular data is deep generative models. Deep generative arXiv:1811.12823v5 [cs.LG] 28 Oct 2020 models have found applications in a wide range of settings, from generating synthetic images [13] and natural language texts [14], to the applications in biomedicine, including the design of DNA sequences [15], and aging research [16]. One important field of application for deep generative models lies in the inverse design of drug compounds [17] for a given functionality (solubility, ease of synthesis, toxicity). Deep learning also found other applications in biomedicine [18, 19], including target identification [20], antibacterial drug discovery [21], and drug repurposing [22, 23].

Part of the success of deep learning in different fields has been driven

by ever-growing availability of large datasets and standard benchmark sets. These sets serve as a common measuring stick for newly developed models and optimization strategies [24, 25]. In the context of organic molecules, MoleculeNet [26] was introduced as a standardized benchmark suite for regression and classification tasks. Brown et al. [27] proposed to evaluate generative models on goal-oriented and distribution learning tasks with a focus on the former. We focus on standardizing metrics and data for the distribution learning problem that we introduce below. In this work, we provide a benchmark suite—Molecular Sets (MOSES)—for molecular generation: a standardized dataset, data preprocessing utilities, evaluation metrics, and molecular generation models. We hope that our platform will serve as a clear and unified testbed for current and future generative models. We illustrate the main components of MOSES in Figure 1.

2 Objetivo general

Construir un marco de trabajo para simular por computadora el diseno de farmacos nuevos, que incluya servidores en la nube, software, procesos, procedimientos

3 Objetivos especificos

1. Instalar configurar los servicios en servidores
2. Desarrollar y/o instalar software de IA para la simulacion insilico
3. Documentar los procesos para realizar las simulaciones
4. Validar el experimento contra otros modelos realizados in vitro, en animales, entre otros.
5. Publicar resultados

4 Cronograma de actividades

1. Investigar el estado del arte del modelo insilico y herramientas existentes (www.insilico.com)

2. Establecer modelo de computo en la nube para la instalacion y configuracion de las herramientas de software
3. Entrenanamiento en herramientas tensorflow y keras basadas en python
4. Entrenamiento en redes neuronales VAN y modelos generativos (ejemplos con rostros, flores, aves, entre otros)
5. Instalacion del modelo de base de datos de moleculas MOSES (5,000,000 de medicamentos)
6. Instalar, configurar modelos existentes como el modelo GENTRL de la empresa china insilico.com.
7. Probar el modelo pharmaio
8. Probar el modelo pharmaio xxx
9. Seleccionar una caso de estudio para simular la generacion de un farmaco
10. Codificar las redes neuronales para conducir el experimento
11. Validar el experimento
12. Documentar procesos y resultados
13. Publicar hallazgos de la investigacion