WASP Software Engineering Course

Assignment 2: Essay about Software Engineering.

Juan Viguera Diez

My research in less than 200 words:

The problem of drug-discovery can be compared to finding a needle in a haystack. Among the $10^{60} - 10^{100}$ theoretically possible drug-like compounds [1], the aim is to find molecules with a set of properties (such as solubility or low toxicity) [2]. Machine learning (ML) has shown promising results for the task of efficiently exploring the vast chemical space [3] searching for candidate compounds. However, most of the proposed methods do not consider the 3D structure of molecules, which strongly influences some of their properties [4], and sampling molecular conformations remains to be a challenging problem. In my PhD, we aim to design a machine learning model that is able to efficiently generate physically realistic 3D structures of drug-like molecules, overcoming limitations from traditional methods such as poor mixing and low acceptance ratio. To do that we aim to use deep generative models in a similar way as they have previously been used to generate realistic images [5] or music [6]. Specifically, we are using Boltzmann Generators [7], which are generative models able to learn from both examples of 3D molecular conformations and from an energy model.

General reflections:

I would like to start this essay by reflecting about automatic testing and data acquisition. Automated testing, as the name indicates, is a technique that tries to design protocols that automatically assess the performance of a piece of software. Therefore, this practice aims to reduce the time invested by humans in checking the right functionality of software products. I think this idea has a great potential in the field of drug-discovery, in which I work. Drug-discovery is the process of finding a molecule that induces a certain biological activity which helps diagnosing, treating, curing, or preventing a disease. In this application it is crucial to make sure the compounds that are offered to the wide public are safe and effective. Computer programs can be used to make predictions about properties of a certain molecule and even to generate molecules that satisfy a set of desired properties [3]. Although exhaustive testing (both in in-vivo and in-vitro models) is performed to new drugs before they hit the market, making sure the models are as effective as possible helps to allocate resources optimally to maximize the probabilities of finding safe and effective drugs. But what does testing mean in this context and what does it have to do with Artificial Intelligence (AI)? In the last years, several approaches to produce molecules with defined properties have been proposed [3]. The most successful works use deep generative models (an instance of AI algorithms) to generate drug-candidates. In this context it necessary to evaluate the performance of the generative models by performing an automatic characterization of the generated molecules by profiling their physicochemical properties. Some examples could be their stability (if the molecules can exist in the real world), toxicity, solubility, synthesizability (if the molecule can be made using other producible compounds), possible energetic configurations (if are they likely to be found at room temperature)… Therefore, automatic testing, consisting of physio-chemical profiling of molecules generated by deep generative models has a lot of potential in the next generation of drug-discovery pipelines. AstraZeneca has publicly released the algorithms they use for the previous mentioned tasks (https://github.com/MolecularAI) and many other pharmaceutical companies are developing similar methods such as Novartis (https://github.com/Novartis).

Testing is not the only feature that can be automated. Data acquisition automation is also regarded as a very promising direction, both for research and production. With automated data acquisition I refer to a technique in which a ML algorithm autonomously "asks for" certain data to improve its performance. In the context of drug-discovery, this usually means performing experiments in the lab to characterize some molecules. This process can also be automated. One promising approach is to use active learning techniques, in which ML algorithms queries data in which their performance is weaker, combined with automated labs gathering that data. We have a research group in Chalmers working in this field (https://www.chalmers.se/en/departments/bio/research/systems-biology/king-lab/Pages/default.aspx) and both tech companies such as Intel and pharma companies such as AstraZeneca have shown interest on it. As we have read in the first pre-reading paper, dataset management is one of the most challenging parts of software engineering for ML. Therefore, the success of this approaches in real-life applications is intrinsically linked to the effectiveness software engineers show to adress the problem of quickly changing datasets and corresponding model updates.

One thing that makes molecules a unique type of data type is that they can be codify at an arbitrary level of complexity. The simplest way to codify a molecule is by a string with information about the connectivity of atoms (SMILES and SMARTS are examples). The advantage of this approach is that one can use all the tools developed in the context of Natural Language Processing (NLP) to work with molecular systems. The next level of complexity is to describe molecules as graphs with node (e.g. atom types) and edge (e.g. bond types) features. Under this representation, the power of Graph Neural Networks can be used. Finally, there is even the possibility codify molecules at a different quantum levels, for example by using Density Functional Theory (DFT) descriptors.

Software Architecture is usually referred to as a "blueprint" of a system in the sense that provides an abstraction to manage the system complexity. Software Design takes this abstraction and makes sure functional requirements are accomplished. Software Architecture and Design in the context of drug discovery has to deal with all these different representations of molecules, which can be challenging. Nowadays, there isn't a software product that deal with molecules at an arbitrary level of complexity and integration and comparison among methods is a very time demanding task. Provably, different representations will need to be treated separately at low level and put together in a high-level framework.

As a last point, I would like to touch upon Regulations and Compliance. Generally, Compliance means conforming some regulations. These can be specified in the shape of a law, standard, policy etc. Data privacy is central in drug-discovery process, especially when related to clinical data. Data must be anonymized and comply with well-defined and usually strict standards such as the European Regulations (GDPR). Moreover, each company has different compliance rules and ethical committees, which makes it even harder to converge to general solutions. However, there are approaches aiming to share data in a privacy-preserving and safe manner. Federated learning is one specially promising direction in which a ML model is trained across different decentralized devices holding a private dataset. The key of this approach is that there is not data exchange between the different devices, but the model is still trained using the data contained in all the private datasets. In the field of drug discovery, there is one European project called MELODY (https://www.melloddy.eu/) that aims to leverage the world largest collection of bioactive small molecules. This is a collaboration between 10 pharmaceutical industries.

Future trends in Software Engineering in my field:

I would like to start this section by acknowledging that the field I work at is changing very fast. Traditionally, professionals working in drug discovery were mostly chemists. Relatively recently, with the fast evolution of computers, the field of cheminformatics appeared. One example of great success in the field were computer simulations of molecular systems. The field slowly started attracting people with specialty in computer science. However, with the avenue of deep learning models, both, chemists felt motivated to learn and about ML, and ML specialists regarded cheminformatics as an interesting application filed of their algorithms. The reality of the field now is a combination of people with very different specialties trying to tackle common problems. Because of this, I think people in the field see ML as a "tool in the toolbox", but not the solution to old the problems. Indeed, I believe that classical methods will still be in the backbone of the next generation of software products and that mixed approaches, exploiting the knowledge coming from both, known chemistry and data, are very promising. One specially interesting example of this is the field of ML potentials, which can produce simulation data with quantum resolution at the computational cost of a classical forcefields. Classical methods will therefore need to be updated, supported, and maintained while products evolve. In that sense, I don't think ML will "eat software" here.

Additionally, the field of drug discovery requires deep understanding of the underlying molecular systems. This expertise is provided by natural scientist, and therefore I don't think "molecular data scientists" will just become a sort of software engineers and vice versa. I think teams will need to stay diverse. Different people will bring different expertise, but it is hard to concentrate all the required expertise in one induvial. The current trend is that teams of natural scientists and specialized data scientist design the solutions and then software engineers transform prove-of-concept models into products.

[1] Gisbert Schneider and Uli Fechner. Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov, 4:649 – 663, 2005.

[2] Sara Romeo Atanace and Juan Viguera Diez. Towards molecular design with desired property profiles and 3D conformer generation using Deep Generative Models. Chalmers Open Digital Repository. https://hdl.handle.net/20.500.12380/302827 .

[3] Thomas Blaschke, Josep Arús-Pous, Hongming Chen, Christian Margreitter, Christian Tyrchan, Ola Engkvist, Kostas Papadopoulos, and Atanas Patronov. Journal of Chemical Information and Modeling. 2020 60 (12), 5918-5922. DOI: 10.1021/acs.jcim.0c00915.

[4] Benjamin Kurt Miller, Mario Geiger, Tess E. Smidt, Frank Noé. Relevance of Rotationally Equivariant Convolutions for Predicting Molecular Properties. https://arxiv.org/abs/2008.08461v4 .

[5] Aliaksandr Siarohin, Enver Sangineto and Stephane Lathuiliere, Nicu Sebe. Deformable GANs for Pose-based Human Image Generation. https://arxiv.org/abs/1801.00055 .

[6] Jean-Pierre Briot, Gaëtan Hadjeres and François Pachet. Deep Learning Techniques for Music Generation - A Survey. http://arxiv.org/abs/1709.01620 .

[7] Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Frank Noé, Simon Olsson, Jonas Köhlerand and Hao Wu. Science, 365 (6457). DOI: 10.1126/science.aaw1147.