HIGH SCHOOL EDITION
Journal of Student Research
Volume 12 Issue 4 (2023)

The Process of AI-Aided Drug Design

Aditya Shirolkar[1], Rajagopal Appavu[2] and Jothsna Kethar[2]

[1]Tesoro High School, Las Flores, CA, USA
[2]Advisor, Gifted Gabber

ABSTRACT

Artificial Intelligence (AI) is a growing field in today's world and plays a part in many industries today. Its role in drug design and the biological sciences has begun to expand in recent years. DeepChem is an open source tool that explores and employs the methods behind drug design. The tool's process and end result will be indicative of how well AI can perform the job of drug discovery and how much it can expedite the process, as well as reveal the future of tools like DeepChem. DeepChem handles everything from data processing to fitting AI models to performing predictions on proposed molecules. By applying DeepChem and reviewing it, we will be revealing AI's power and limitations in biomedical chemistry and technology. In addition, other AI tools, such as Chemistry 42 and inClinico by Insilico that achieve other parts of the drug design process. Completing a comprehensive review of these methods will provide an overview of what can be improved and the scope of AI as a big money-maker and solution in the biomedical field. The synthesis of drugs is a complicated process that can be simplified by AI tools. This paper explores how AI tools operate and their limitations in the medicinal world.

Introduction

Artificial Intelligence (AI) is a growing field in today's world and plays a part in many industries today. Using many different machine-learning methods, AI has grown to be an accurate and powerful tool for almost anyone to wield. AI chess bots have already begun to become almost impossible to beat, like Stockfish. The hope is that AI can grow to be that powerful in almost any field and aid in accomplishing a lot more. AI-aided drug design is another application of AI that has given rise to a number of tools used in the field of medicine. In 2023, Insilico, an AI-based pharmaceutical company, announced that its drug, whose design was aided by AI, had entered human clinical trials in just under 30 months, which is a rapid feat. AI-aided drug design is the route of the future, and uncovering how it works is crucial. A deep dive into DeepChem, a Python library, will yield information and potential routes for how to improve AI applications. As with any AI-related breakthrough, the limitations must also be highlighted and discussed.

History of Drug Design

In general, drug design over the years has expanded rapidly, from the synthesis of chloral hydrate to the development of acetylsalicylic acid, better known as Aspirin, to more recent movements involving AI. The first medicinal drugs came from plants, herbs, and fungi. Chloral hydrate, a sedative-hypnotic still found in some countries today (Jones, 2011, abstract), was discovered in 1869, as were other drugs integral to surviving certain diseases today. Following this development, penicillin was perfected around World War II, allowing for treatment of wounds and keeping them from being infected. Taxol^{R}, a drug that treats certain cancers, hit the shelves in 1994 after earning FDA approval in 1992. It was the synthesis of the drug that signaled a change in medication development. As AI took hold in the 2010s, its integration into the medical sciences slowly took hold. And as data surrounding established drugs was recorded and stored, AI became a realistic tool for drug design. Nowadays, the process of drug design involves a lot of time and technology. The hope is that AI can cut down the amount of time it takes to get a tested medication into the market.

AI can help with target discovery, drug synthesis, and even in predicting the results of clinical trials. Drugs used to take around 5 years to develop and years more to test before releasing the drug into the market. But with AI, the entire process could take as little as 2 to 3 years to immediately be released into the market.

Artificial Intelligence Overview

Artificial Intelligence (AI) refers to a machine's ability to learn like a human. After the development of computers from the 1940s to the 1970s, computers became a valid path of data processing and had continually improving processing times and results. These advances in computing power allowed machines to start to make decisions, and in 1997 IBM's chess bot Deep Blue was able to defeat world chess champion Gary Kasparov. Deep Blue had to evaluate every possible move and make its decisions based on that, unlike the AI models today which use related data to draw inferences. After 2010, the processing power of machines took a huge leap forward, allowing machine learning to happen much faster and abide by more constraints. Combined with the vast amounts of data available during this time, AI tools began to take off as companies introduced image processing and even voice processing. AI's role in medicine could be crucial in time, potentially helping to find cures for rampant diseases. Figure 1 below shows the skyrocketing attention the AI market has been gaining in recent years, and its potential for huge breakthroughs in the coming years in every field.

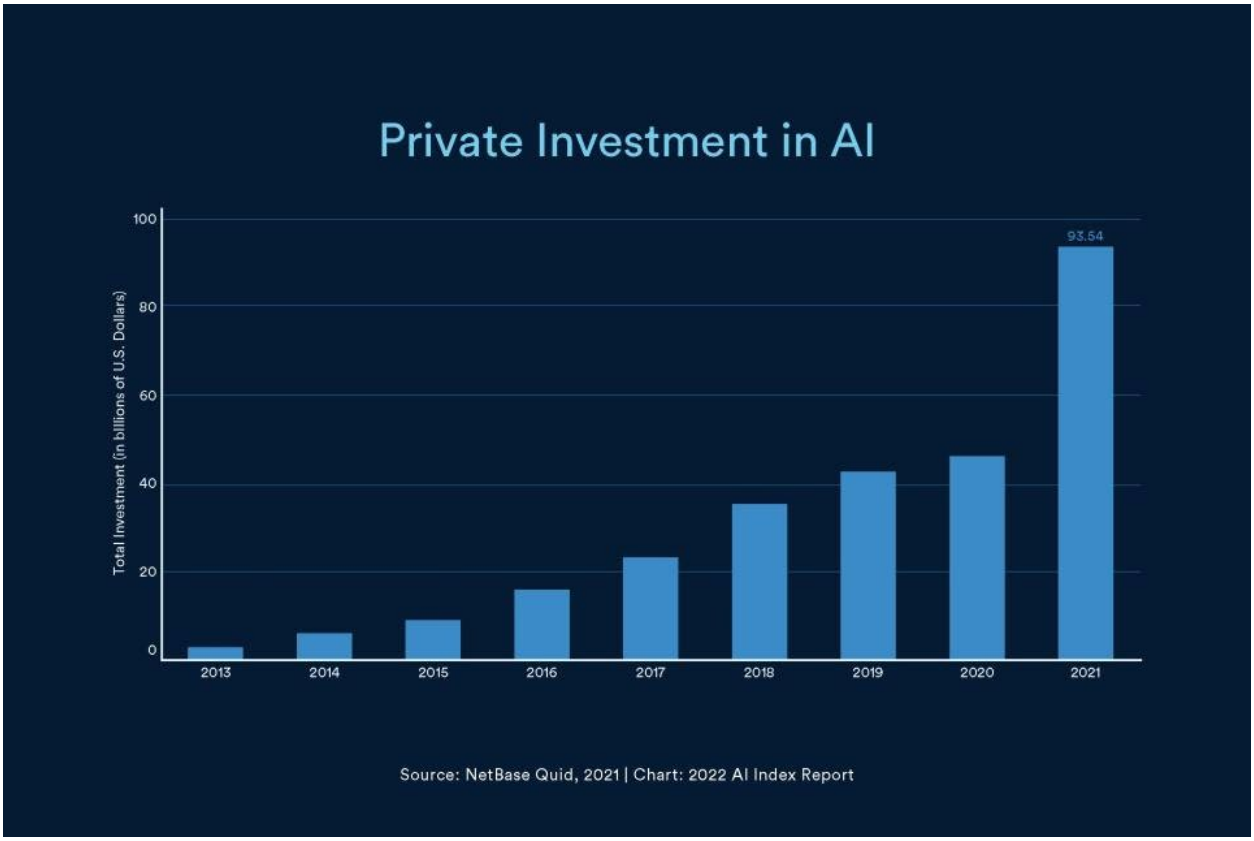

Figure 1. Growth of investment in AI by 2013-2021. Created by Stanford University Human-Centered Artificial Intelligence. Data from NetBase Quid 2021.

AI tools in the biomedical industry are relatively new, since they involve huge sets of data to make predictions off of. Relevant data was unavailable for the industry until around 2014, but with the data models can function. Improving these models relies on bigger and better data. Luckily, DeepChem has these datasets ready to use and processed for models. In addition, since DeepChem is open source on GitHub, the user has the ability to add their own experimental datasets to produce and train models on.

Data Conversion for Machine Learning

SMILES Conversion and Usage

In order to train an AI model to suggest new drugs, the data must be converted into a manner that the machine can interpret and connect to other variables. Most variables, like solubility for example, can be quantified and are therefore easier for the machine to connect. However, molecular structures are not quantifiable and must be converted into a simpler manner for the model to connect structure with solubility or another variable, such as toxicity. This is where the SMILES method comes into play. SMILES (Simplified Molecular Input Line Entry System) simplifies molecular structures into a set of rules that allow the model the necessary information to connect aspects of a certain drug to that of the target drug. All elements from the periodic table are present in the SMILES system, but most organic compounds center around carbon, nitrogen, oxygen, and hydrogen. Some SMILES conversions are given in Figure 1 below. The first rule of SMILES conversion is that either no hydrogen bonds need to be represented, or all hydrogen bonds need to be represented. In the examples below, no hydrogen bonds were represented as this is the method used in the DeepChem library (explored later). For longer molecules with carbon chains and extensions in the structures, extensions (or branches of the molecule) are given in parentheses, along with possible extra bond notations. 2-Propanone, the example given by the EPA (Environmental Protection Agency), would be notated are CC(=O)C. The second carbon in the chain has a branch double bonded (notated by =) to an oxygen. The SMILES system automatically infers hydrogens to fill in bonds for the structures, so no other information is needed. Online, there are a number of tools that will convert SMILES structure to molecule names. When using DeepChem, I highly recommend using one of these tools to see what molecules are in the datasets. The one utilized in this research was called Syntelly and is linked in References below.

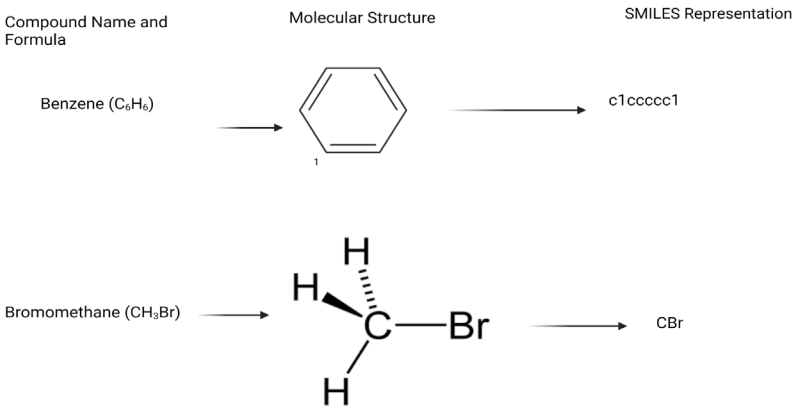

Figure 2. SMILES Conversion of simple molecules.

Other Data Conversions

SMILES is just one way to get machines to understand one of the metrics behind drug development. Other data must also be converted for AI to work effectively at modeling new target drugs. The DeepChem library has all these converted datasets so that all that is left to do is apply the AI models to the sorted datasets. This will be an overview of the data conversions the DeepChem developers have made, while the full datasets themselves can be found online for further reading. The Clintox dataset has data containing the SMILES structure of molecules and whether or not the substance passed FDA tests for toxicity. A passed toxicity test is a one in the column, while a failed toxicity test is a zero. Next, the Delaney dataset contains SMILES structures and the solubility, measured in log(molarity). By converting the extremely small molarities (the complex structure of the molecules leads to a high molar mass, hence the miniscule solubilities) into negative numbers, AI models can better work on fitting a trend between structure and solubility. Solubility is an important component in drug discovery because if a proposed drug isn't soluble enough, it will not have the intended effect on the patient. The PDBBIND dataset contains SMILES structures and binding affinity for protein-ligand interactions, which helps predict how a proposed molecule might bind with a given protein. Each molecule has been given a one if it can favorably bind with a given protein and a zero if it cannot. By converting the data into binary, it opens the route for AI to predict properties of a new proposed molecule.

Predicting Solubility with DeepChem

As discussed before, DeepChem's Delaney dataset contains data on solubilities and molecular structures. In Python, the data must be loaded into an object and split into training and test datasets. The online tutorials show how to start off by loading DeepChem and loading the dataset. Next, DeepChem's provided GraphConv model is used in the tutorial, although for this research, we will play around with some of the settings to try to improve on the accuracy of the models. For this section, we will stick to the settings in the tutorial without alterations. Next, once the datasets have been loaded and split into training and test sets, all that's left to do is fit the model and use the predict feature to test it. In order to measure the accuracy of the model, let's use the Pearson correlation (r^2). Based on the default settings from the tutorial, and in almost every test we did, the model had a better performance on the training set than the test set. This is known as "overfitting", in which models perform better on data they have seen before than unknown but similar data. Once a working model is accomplished, eliminating overfitting while maintaining the accuracy of the model is the main goal. A score of one for the model would indicate a perfect prediction, while a score or zero would indicate no perfect predictions at all. The scores can therefore be considered almost like percentages; closer to 1 would indicate a better performing model and vice versa.

Alterations	Training Accuracy	Test Accuracy
Default, Layers (128,128), Dense layer size 128, Epochs 100	0.92	0.65
Layers (256), Dense layer Size 256	0.96	0.68
Layers (100), Dense layer Size 256, Epochs 100	0.97	0.74
Layers (256), Dense layer Size 300, Epochs 1500	0.98	0.75

Figure 3. Accuracy of model in predicting solubilities based on alterations.

In DeepChem, each individual layer can be defined with as many neurons as desired, but increasing the number of layers often led to a lower accuracy rate for both training and test accuracy. Increasing the number of neurons seemed to do the trick though, as the model could process more information. The dense layer size is a neural layer in which every neuron takes in an input from the previous layer, as opposed to having streams of neurons. Whenever the dense layer size was equal to or greater than the single layer defined, it led to a generally better accuracy since every neuron is processing all the data and producing outputs that are later averaged out. Finally, the epochs refer to training epochs, or how many times the model trains with the dataset. Increasing the number of epochs also generally increased the accuracy of the model, but it also increased the time the model needed to fit, so increasing the number of epochs should be done slowly. For reference, the default settings of the model were two layers with size 128, a dense layer size of 128, and 100 epochs. By increasing the number of epochs substantially, the model became much more familiar with the training data and was able to better apply its training to the test data and obtain a decent accuracy.

Predicting Protein Binding Affinity with DeepChem

This section will focus on how the AI model predicts binding affinity of a protein-ligand complex. Tinkering with the AI model from the DeepChem tutorials leads to error messages, so read the whole DeepChem documentation and then proceed to load models. The PDBBIND dataset (see Other Data Conversions) contains SMILES structures and the names of proteins that the structures are proven to bind to. DeepChem's complex tools along with RDKit packages allow DeepChem to visualize how ligands can slot into protein pockets. Figure 4 below demonstrates how ligands, or drug molecules, slot into proteins to be consumed and produce therapeutic effects.

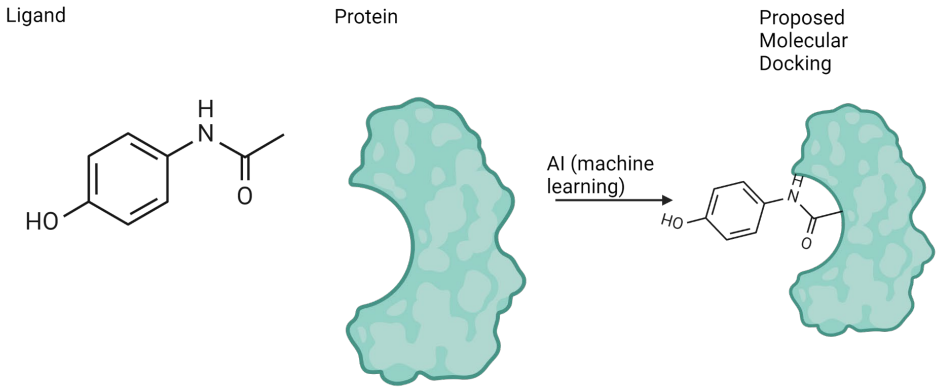

Figure 4. Protein and Ligand and Proposed Position of Molecular Docking.

Next, by invoking an Sci-KitLearn model and feeding it the featurized PDBBind dataset, we were able to get a working model. But when using the Pearson squared coefficient to determine its effectiveness, the model has a respectable 0.88 on the training data set but only a 0.0077 on the test data set, which means the model is largely ineffective on similar data. Possible solutions could include scaling down the dataset to train the model based on connections between ligands. For example, some ligands contained carbon rings but had different binding affinities to the same proteins, so by scaling down the dataset to include only these it could be possible to get more accurate predictions.

Predicting Toxicity with DeepChem

Predicting toxicity is a relatively straightforward process thanks to DeepChem's documentation. By loading the Clintox dataset, which contains SMILES structures and failed or passed clinical trials (see Other Data Conversions in Data Conversions for Machine Learning). By using a simple GraphConv model, each structure is classified as pass or no pass for a clinical trial. Altering the layers and epochs yields different Pearson Correlation Coefficients every time. Figure 5 below demonstrates the training and test Pearson Coefficients based on the alterations.

Alteration	Training Accuracy	Test Accuracy
Default, Layer Size (128,128)	0.92	0.56
Layer Size (256,256)	0.95	0.64
Layer Size (128,128), Epochs 150	0.94	0.62
Layer Size (256,256,256), Epochs 175	0.97	0.73

Figure 5. Recorded accuracy of the GraphConv model on toxicity data.

In general, increasing the size of the layers and the number of epochs both led to an increased accuracy. For this model, we did not experiment with an incredibly large number of epochs, but the increased number of epochs did show evidence of reducing overfitting. However, since protein binding affinities are given in numbers (and so are solubilities), the values predicted by the model could have a huge range. This is further explained in the section Challenges of DeepChem.

Combining the results of the models

The ideal model would be able to combine all three of the previous parameters (and others if desired) and produce a resourceful chart or CSV file that can be compared with the true values for accuracy. Any model that is going to have a scope in drug design needs to be able to predict all these properties simultaneously in order to lessen the amount of time it takes to develop a new drug. While DeepChem does not have this function yet, and its models are not accurate enough, bringing this feature to DeepChem could be a revolutionary advance forward. There are some protected company tools (see Other AI Tools and their Uses below) that can perform this function.

Other AI Tools and their Uses

DeepChem is a great tool, and its open source, which is why it was chosen for this paper. But there are other tools out there that incorporate AI and are in use, so this section is merely for discussing only those tools. First off is Insilico's entire trifecta of AI tools, named PandaOmics, Chemistry 42, and inClinico, offer support for every part of the drug design process. PandaOmics uses -omics (genomics, phenomics, proteomics, etc.) data to come up with target discovery based on disease genomics. The idea is that genes of a disease are analyzed, and related diseases based on genes and their cures are weaved together to produce a target drug that can be built. In addition, PandaOmics predicts the chances of a potential target to enter Phase 1 clinical trials in the next five years (Insilico, 2023, PandaOmics for Research). Once a target has been identified, Chemistry42 perfects the design based on desired properties, such as minimal toxicity (if any), increased solubility, and optimized binding affinity (too high of a binding affinity could

mean the drug gets consumed too quickly by the proteins). As the drug is generated and tested, its properties are mapped and given to possible vendors for similarity and novelty. In this manner, new drugs will automatically be added to existing datasets to lead to even more possibilities for drug discovery. Once a new drug has been synthesized and mapped, its properties go to inClinico, which predicts the results of the drug in clinical risk assessments. Once again, it makes these predictions based on similar drugs and how well those drugs performed in trials. Of course, the drug is still tested in these trials, but it makes the process much easier knowing the outcomes are likely to favor a certain outcome. Insilico's newly developed drug, called INS018_055, followed this process with the AI tools and entered Phase 2 IPF trials in June 2023.

Another AI tool that deals with drug discovery and design is AIDDISON. This tool also includes much of the features of DeepChem, including de novo drug design and molecular docking predictions. The software is protected (to avoid plagiarism) and offers a time-saving and cost-efficient optimization for drug design.

Challenges of DeepChem

DeepChem's models work decently well, but the problem with AI models is that measuring the loss is usually difficult. For example, if the model gives a result of 5.272, is the actual answer in between 5.271 or 5.273? It could even be between 2 and 8 for all we know. In other fields, this ambiguity might be acceptable, but in medicine results need to be as accurate as possible. Luckily, DeepChem makes it easy to measure the loss and the uncertainty of the model for each individual output. The Pearson coefficient from before is a measure of how the model performs across a huge dataset, but the loss measurement depicts how the model performs for a single synthesized output.

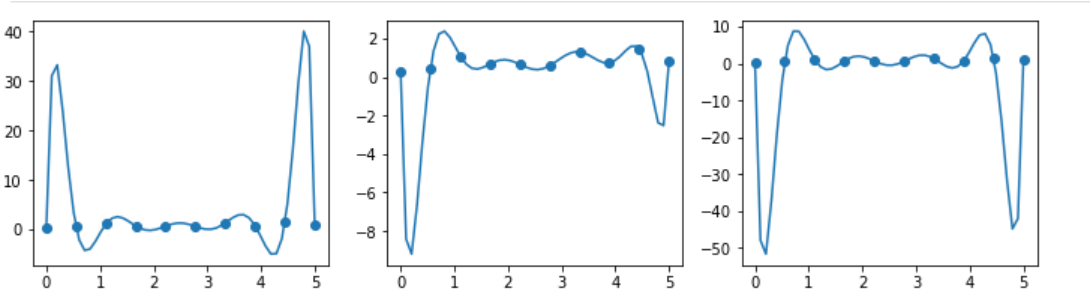

Figure 6. Demonstration of how epistemic uncertainty works in DeepChem. Created and copyrighted by DeepChem: Uncertainty in Deep Learning.

Epistemic uncertainty is uncertainty regarding how the model can fit together a specific number of data points. In calculus, this kind of fitting would be much like a power series, in which each turning point added leads to another term for the polynomial. The examples in Figure 6 use tenth-degree polynomials to fit ten data points into certain curves. However, all ten data points are exactly the same across all ten figures, and the lines represent different ways the model could extrapolate connections. Because of this, DeepChem models tend to lose a lot of accuracy when using them on data that the model was not trained on, as seen in the accuracy data from previous sections.

Conclusion

In general, AI looks like a very plausible tool for drug design. Although its accuracy with new data might be suspect, better AI tools are always becoming available which could lead to a process in which AI handles everything about drug design, from discovery to production to clinical trials. However, with this power comes questions as well. For example, if the drug fails clinical trials, who's to blame: the scientist or the machine? Some food for thought. AI-aided drug design is a very deep field that will require much more learning to completely uncover, but for the basic scientist,

uncovering DeepChem and its properties is a great preview. DeepChem contains many more tutorials for readers to explore on their own, and playing around with these tutorials and uncovering how medicinal AI works will make any person seem like a qualified expert on the topic. By completing this research, we uncovered how to use DeepChem in predicting molecular properties such as solubility, protein binding, and toxicity, as well as uncovering the uncertainty of the models we trained.

References

Chemistry42 | Insilico Medicine. (n.d.). https://insilico.com/chemistry42

Council of Europe. (n.d.). *History of Artificial intelligence - Artificial intelligence - www.coe.int*. Artificial Intelligence. https://www.coe.int/en/web/artificial-intelligence/history-of-ai

From start to phase 1 in 30 months | Insilico Medicine. (n.d.). https://insilico.com/phase1

InClinico | Insilico Medicine. (n.d.). https://insilico.com/inclinico

Jones, A. W. (2011). Early drug discovery and the rise of pharmaceutical chemistry. *Drug Testing and Analysis*, *3*(6), 337–344. https://doi.org/10.1002/dta.301

Linuxize. (2020, June 26). How to Unzip (Open) GZ File. *Linuxize*. https://linuxize.com/how-to-unzip-gz-file/

Lynch, S. & Stanford University. (2016, March). *The state of AI in 9 charts*. Stanford HAI. Retrieved August 3, 2023, from https://hai.stanford.edu/news/state-ai-9-charts

MoleculeNet — deepchem 2.7.2.dev documentation. (n.d.). https://deepchem.readthedocs.io/en/latest/api_reference/moleculenet.html

Nicolaou, K. C., & Montagnon, T. (2008). *Molecules that changed the world*. Wiley-VCH.

Office of the Commissioner. (2018b). The drug development process. *U.S. Food and Drug Administration.* https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process

PandaOmics | Insilico Medicine. (n.d.). https://insilico.com/pandaomics#rec236934966

Ramsundar, B., Eastman, P., Walters, P., & Pande, V. (2019). *Deep learning for the life sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media.

Ramsundar, B. & DeepChem. (2021). *The Basic Tools of the Deep Life Sciences*. GitHub. Retrieved July 30, 2023, from https://github.com/deepchem/deepchem/blob/master/examples/tutorials/The_Basic_Tools_of_the_Deep_Life_Sciences.ipynb

SMILES Tutorial | Research | US EPA. (n.d.). https://archive.epa.gov/med/med_archive_03/web/html/smiles.html

Syntelly. (n.d.-b). https://app.syntelly.com/smiles2iupac