Graph Neural Networks for drug discovery

Mahesh Bhume
School of engineering and applied science
Aston University
Birmingham
United Kingdom
maheshb64@gmail.com
May,2022

Abstract:

Drug discovery is a time-consuming process it takes more than ten years for pharmaceutical researcher to identify new drug from millions of drug candidates, so the average time and cost for designing a drug would result into more than a decade, it's the reason why machine learning is the ideal tool to solve this problem and to improve pharmaceutical researcher to find drugs quickly in order to respond to recent pandemics or novel pathogens. In a computational terms drug discovery is a search problem, we can put this as, chemical space as sort of maze and ideal drug would kill the virus or pathogen, this is sort of the whole central goal of this drug discovery, but this computational task is challenging, it comes from the size of the chemical space, the number of drugs like molecules is estimated to be around 10^{60} and it's difficult to screen all of them even in pharmaceutical industries standard experimental (high throughput screening acids can do around 10^5, therefore instead of doing brute force search, an efficient algorithms to design automatically and that's where deep learning comes in to solve this problem. This paper discuss how GNNs are useful for drug discovery.

Introduction:

In the last fifty years discovery has enabled drugs that have taken diseases that used to be death sentence into something that is now manageable or curable that includes vaccines for infections, diseases, cancer immunotherapies, auto immune system modulators it is a remarkable tale of achievement, while on the other side the price tag of approval for a new drug is more expensive, that is not because the single drug incurs that hype in price from its journey from idea to approval is because that one drug carries the previous work on that experiment did not make it. So in the development of drug discovery there is almost 12-15 year journey from idea to approval with many forks in the road, where one fork if successful takes us to 99 don't work and we have limited tools to make a decision on which of the paths in front of us are going to lead us to success and taking the wrong one can be matter of years and the cost with that to understand we took the wrong turn, so how can we make better predictions of downstream outcomes hopefully helping bend the curve. In that event, the machine learning and data science could potentially play huge role in drug discovery.

An important milestone for the betterment of deep learning is ImageNet, which was created at Sandford by Professor Fei-Fei Li

this give machine learning enough training data to the point we could start answering much harder questions such as multi-class classification, object class (feature) in the image and the correct label for these images, today the performance of machine learning for this problem is demonstrably beyond human level performance for tasks that each of human has been trained to perform since infant, so the fact that we could achieve beyond human level is remarkable.

To understand what made this possible would be, more data enables rich models to achieve higher performance and also because the models we use to employ in early days of machine learning, these were usually simple models constructed on the top of human constructed features, by doing feature engineering to feed those algorithms and it has been observed those models were good because they introduce a lot of human bias knowledge but they asymptote it at a relatively low level, today the computers starts out with scratch and just raw features and it takes longer in some ways, it require more data to reach high level of performance but it turns out computers actually end up constructing better features than people do, which is why performance keeps going up.

Today with deep learning the computers inferences with images, starts to get raw features and it construct an increasingly complex hierarchy of features which are built on top of other features, so it's able to construct features that are very subtle and can make distinctions between features. The other aspect of this is new representation learning aspect, so if the original image is 1000*1000 then they said in million-dimensional space and compresses it into hundred dimensions and those are meaningful features of the computers constructed.

So, from a mathematical perspective the machine has created a hundred-dimensional manifold and a million-dimensional space and has embedded all those million-dimensional images in that hundred-dimensional representation.

The deep learning has discovered an antibiotic that can treat previously untreatable strains and bacteria, The antibiotic is a chemical named Halisin, and it was discovered by a pioneering machine learning approach called Graph Neural Networks. The fairly straightforward approach can also be used to find new drugs for other diseases like cancer, improve transportation.

History of antibiotic Discovery:

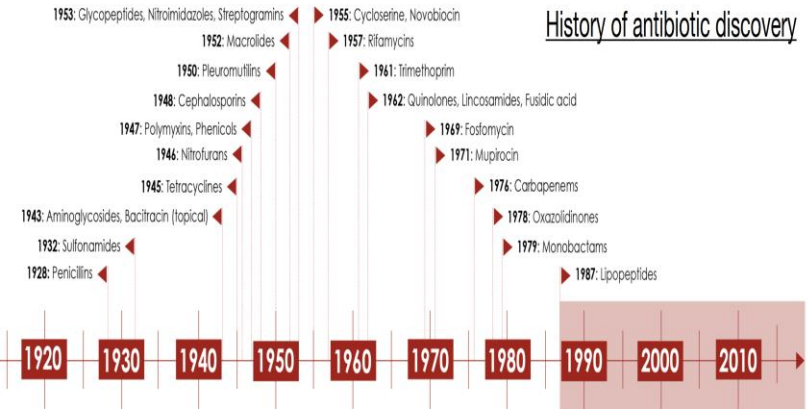


Figure source [7]

Since 1930s we discovered penicillin and it's has been quite fruitful over many decades but after the low hanging fruits were gone, it become very hard discover noble antibiotics. On the other hand, because of abuse of these antibiotics a lot of bacteria's became resistant to the antibiotic in the clinics and that's why it became a huge problem of bacterial infections and there no cure of antibiotic resistant. So, we need to find novel antibiotics to fight against these resistant bacteria's.

Previous work in discovering drug using traditional techniques:

The pharmaceutical researcher community has been looking for different kinds of solutions all based on computation, there are two important concepts one is called functional space, which is basically what's the property of each compound with the mapping from compounds to the functional properties like toxicity that's classic property, which researchers don't want in the drugs, there's also solubility constraints, these properties are measured by numerical score, example would be if toxicity is around 10 then it's not good and if the solubility is equal to five then its good.

Second concept is chemical space, it's basically huge set of potential molecules, the goal is to find chemical structure that has particular set of properties. In the past there were three different types of approaches to find drugs with specific properties.

Three schemes of drug discovery

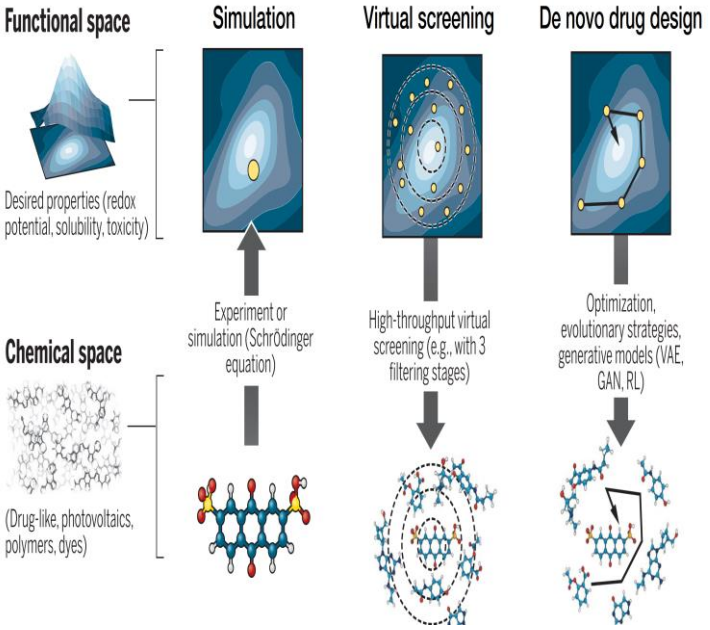


Figure source: Sanchez-Lengeling et al., Science 361, 360-365(2018)

First approach is based on simulation, this could be molecular dynamics or molecular docking, example of this would be how the molecule wiggles around in a protein pocket and by using molecular dynamics we could know the molecular would binds this protein. The simulation is often too slow because it takes days for very accurate stimulation even for just one compound, so researchers have focused on the second and third approach.

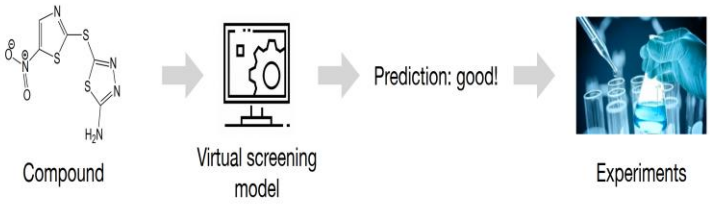
Second approach is called virtual screening, here researchers would instead of doing simulations, they directly try to predict its solubility or toxicity given the structure of the molecule. This approach allows researchers to put these ideal candidates into virtual screen model and then select the top ones that has the best solubility or the

lowest toxicity. Unfortunately, would this approach still researchers would be stuck this the primary problem of not able to put all the 10^{60} compounds into model that will still take forever.

The last approach is called De novo drug design, it's the most ambitious of three, which basically want to solve this inverse problem by looking at the set of criteria's, like low toxicity high solubility and then do reverse engineer (optimization using backpropagation, evolutionary strategies, generative models (GAN, VAE, RL)) the right compound from the set of criteria.

Virtual screening:

Figure source [7]



The input to this model is a compound and then it can predict numerical score, whether it is good or not and that depends on what is the property, like if it's binding affinity then it needs to predict weather this compound has a low binding affinity or if it's toxicity property. Virtual screening is favourable because it is much faster than experimental screening in labs, as it can test around 10^8 compounds within a day while experimental screening will take more than a year and it is quite cheap as its computational rather than in web labs. Virtual screening limits itself within the set of commercially available compounds for example the like Zinc library, basically we put all these compounds in our actual model and then this virtual screening method gives us a rank over these candidates and then we just test the top once in web labs. So, the advantage is there's no need for to synthesis of any compounds, after we generate the ranking the test is relatively easy, as we don't need to synthesise them, but it loses the coverage because it still not comparable to the entire chemical space. These traditional virtual screening techniques are based on hand crafted features, which its accuracy, so in a way the model often produces wrong predictions.

De novo design:

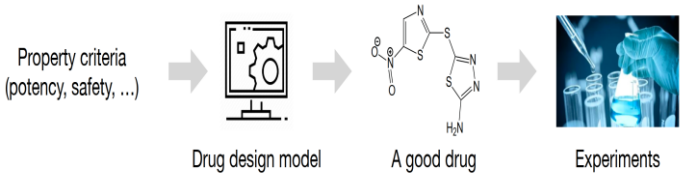


Figure source [7]

It's the inverse to the virtual screening, basically from a set of criteria like potency, safety reverse engineer right compound and that's challenging problem as molecules structure and machine learning models output is just numerical number rather than some sort of structure, so this problem is much harder than virtual screening. The De novo design has trade-off between ease for synthesis versus converge, it can explore the entire chemical space very efficiently because we don't need to try every compound, the model will figure out that for us, but the model will actually suggest some new compounds that's not commercially available, so we actually need synthetic chemist to figure out how to synthesis that compound in order to test them in web labs. So that slows down the experiment workflow, but it gives us more potential in discovering new drug.

The De novo design and virtual screening idea has been there for over two decades but it's quite hard to make them work, as this problem traditionally based on these genetic algorithms by doing random mutation on the molecules and then these mutations are certain time designed rules, so in a way we don't have yet efficient

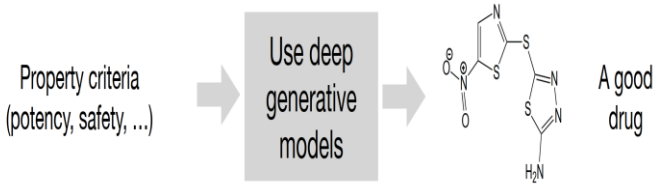
algorithms to explore this chemical space efficiently.

Deep learning: revolutionize drug discovery

These traditional techniques are not efficient and because the rise of deep learning, computer vision has achieved human level accuracy and the key to success is achromatic feature learning, and contrast to this in virtual screening or traditional methods were based on handcrafted features, which inherently limits the performance.

Deep generated models can help us to generate a compound with desired properties.

Figure source [7]



This no longer requires handcrafted designed rules for how to mutate a molecule, so the model can learn how to generate them and cater them with respect to specific properties, so that's why deep learning revolutionize drug discovery field and in deep learning especially graph neural networks algorithm.

An example of what these can do, and the answer would be a discovery of powerful antibiotic called Halisin, which can kill lot of bacteria's which are resistant to existing antibiotic. So, this an example what we can achieve using these GNNs that previous methods cannot achieve.

Figure source [7]

What is Graph Neural Network

Graph is a structured way of representing data, it consists of nodes to represent entities and edges to represent the relationship between these entities. Each node has set of features which describes the input data and edges either could be undirected – meaning two way and directed – one way. Extending from undirected to directed graphs is not difficult once we get the basic mechanism.

Why Graph Neural Networks for this application: As we have seen expensive success of Neural Networks had on many problems, they are able to learn variety of desired outputs from serval types of inputs such as numbers like stock prices, words that we use every day or even images like, so don't we feed drug discovery data through the all-powerful neural network, it turns out that neural networks have mostly been successful on more structured data types. For example, numbers belong in a series, words that we use every day are found in sentences, images as they come as a grid of pixels. But neural networks still struggling with more complex data such as graphs, which do not have any fixed structure or ordering. So, in order to reasons with graphs, researchers went back to the drawing board to study what makes neural networks so successful. The first layer in a common convolutional neural network takes each pixel and extract information about the region by aggregating or selecting maximum pixel within the context of its

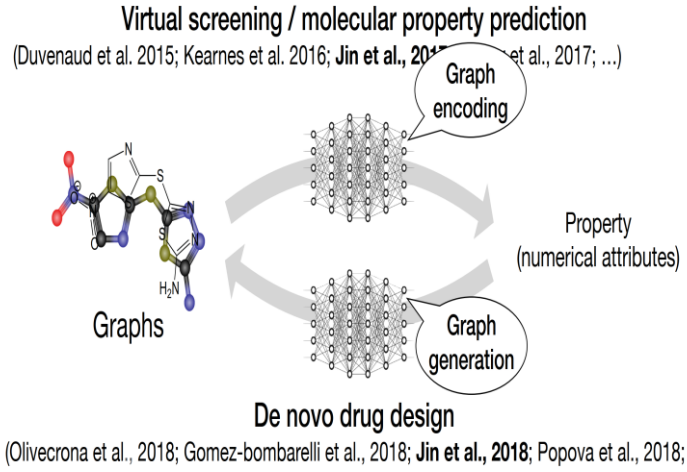
neighbourhood through stride (sliding window technique), the next layer then extracts information about that region within the context of its neighbouring regions. Doing this over enough layers, the network will be able to reason over different parts of the whole image and then with some liner computations, it can finally identify the relevant object. To apply this mechanism to graphs, if we look closer at an image, we can view it as special type of graph, a grid graph that is structured and every node is a pixel that is connected to its neighbouring pixels, so if we use same aggregating or maxing mechanism, we can actually use that same intuition on graphs. Relaxing the properties of fixed structure and ordering, we can similarly gain a better understanding of each node by aggregating the message received from all neighbouring nodes. Over sufficiently rounds of message passing, we will eventually obtain a final representation of each node in a graph that well describe it gives the larger context. We can also view the rounds of message passing as a series of layers just like in our convolutional neural network. Here we go from the input to layer one of message passing and then layer two and this produces our final understanding of nodes.

In the process of reasoning over a graph, the graph neural network seeks to summarize all the information of each node into a numerical representation, where nodes which are more similar will be closer to each other. We call this representation of nodes as node embeddings, and we call the space of all possible representations as the embedding space. But to come up with these embeddings and where to place them in the embedding space, neural networks need an appropriate measurable objective to guide the network into learning something useful for our task. So, rearranging the expressions, we can quantify our objective into a loss function to tell the network how well it is doing. Adding in all the other objectives from other nodes and taking the average, we now have a working loss function.

Firstly, from the input, we forward propagate through the layers to get a final representation of each node. We then tabulate the loss and back-propagate it through the network to tell it where it went wrong. The network changes its computations and tries again for better performance, we can do it as many times as we want and after the network has learned embeddings of graph that are good enough to meet our objective, we could simply shortlist few compounds, which are in a certain radius or cluster, filter out some compounds which could be deal breakers, such as higher toxicity. This operation is a form of link prediction, where we task the network to recommend new potential edges between nodes. Another popular approach is node classification, where we use the representation of each node to predict a certain class. We also have clustering, where we break the embedding space into different groups. And finally, we have graph classification, where we do some form of aggregation over all node representation of a graph and compare it with other graphs. We can notice that the key component of these strategies involves the need to learn good node embeddings.

Main technique: Graph Neural Networks (GNNs)

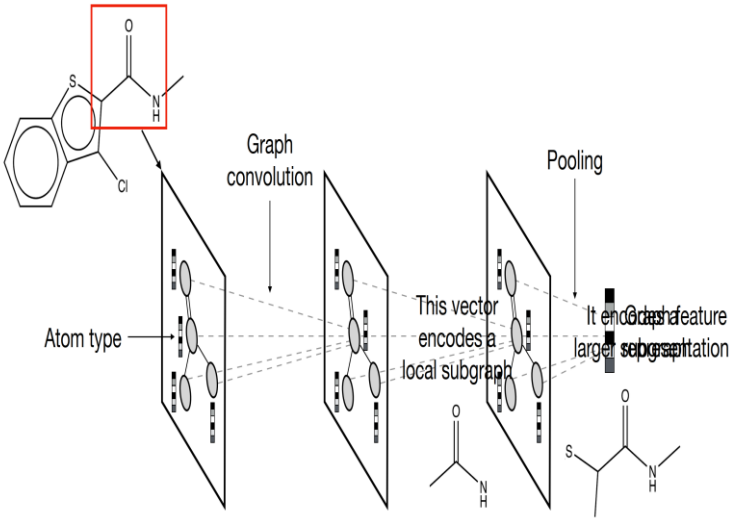
Figure source [7]



If we look at these two problems, this virtual screening also known as molecular property prediction and also the De novo design, it basically this loop right from the molecule to the property and from property to the structure and reason why GNN becomes useful, as molecules can be modelled as graphs. For example, each atom in this molecule can a node in this graph and each bound is the edge in this graph. Now the problem like virtual screening is learning how to encode this graph into some sort of representation then based on that representation predict the right property and in the De novo design, the problem is graph generation task, given a set of criteria generate the graph using the deep generative models.

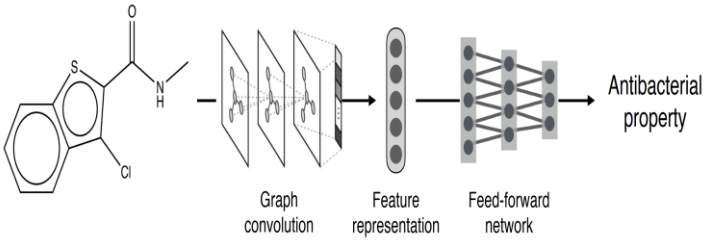
How Graph Neural Networks works with discovering drugs:

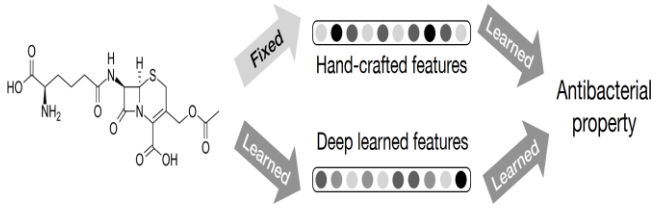
Figure source [7]



GNN is standard deep learning model input to the model is a graph, so we need to represent each molecule as a graph where each atom is a node and each bound is an edge. Given this graph structure we need to convert it into a continuous feature vector, as none of the models can directly operate on graphs, this procedure is called graph convolution. Initially each atom has a vector that indicates its atom type, whether it is nitrogen or carbon. Now after having these vectors, we can apply graph convolution, it similar to convolution in images, just squash these vectors into a single vector by a linear layer, followed by some non-linear activation like relu or sigmoid. So, the squashed vector encodes a local sub graph, looks like above figure. The good thing about GNNs is it's all continuous, so we are not constructing long high dimensional feature vector but rather we're condensing everything into this low dimensional continuous space, and this allows feature now is low dimensional we

can actually generalize better to new dataset and to new molecules. At the end we apply pooling operation that combines all these vectors into single vector that summarize what this graph looks like, then add a feed forward network on top of this, feature representation learned by this GNN to predict a single numerical score, basically a probability of how likely it can kill the bacteria.





Pulling back this paradigm traditional approach use this handcrafted feature and then learn something on top of this feature vector but now we have these two steps learned jointly within this GNN architecture.

Comparing GNN with other deep learning models:

Figure source [7]

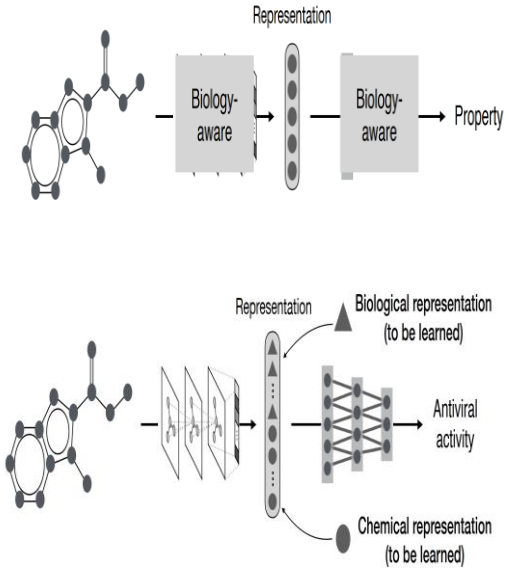
So far, we've seen GNNs are better than traditional methods, can models also discovers halison and it turn out that for example feed-forward neural network based on Morgan fingerprint it does perform bad, as much as the rank of halison is around one thousand, where it tested around only 100 compounds were passed, so in a none of these traditional methods even some of them using neural networks cannot discover halison. In a way it shows

Model	Feature	Rank of Halicin
Graph neural network	**Learned**	**61**
Feed-forward neural network	RDKit features (fixed)	273
Feed-forward neural network	Morgan fingerprint (fixed)	1217
Random forest	Morgan fingerprint (fixed)	2640
Support vector machine	Morgan fingerprint (fixed)	771

that these learned features are better than these hand designed features.

Incorporating biological knowledge into GNNs: The previous model is just looking at the structure by itself and the input is just molecular structure, and we are sort of inferring some biological knowledge just from data. So, in a way this model is losing a lot of information, properties may depend on likes of additional biological information. For example, whether this molecule binds to certain proteins or biological targets and maybe that's why it killed the bacteria, in a way this model does not tell us that much about the inner mechanism or how it actually arrives at the end property. So, we need to bring back this biological knowledge into the model in order to generalize better. Hence, we need to build a social model that are aware of certain biological knowledge.

Figure source [7]



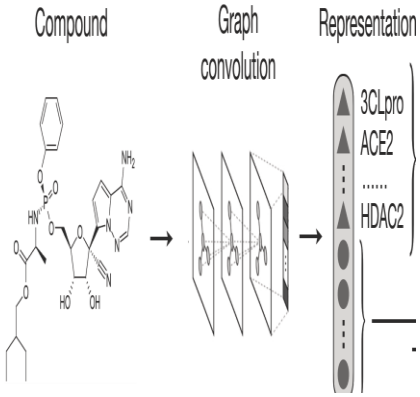
That lead to our next model called ComboNet, which try to incorporate both biology and chemistry. The difference between ComboNet and the standard GNNs is in the molecular representation. Before each molecule just learned these fingerprints that's purely just encoding the chemical structure of the molecule, can be seen in the following figure.

Now we can add additional knowledge like what kind of biological targets that this molecule inhibits, and these are represented by these triangles, that in contrast to the previous chemical representation is shown by these circles. Now we can put both of these representations into the model to predict the antiviral activity of a drug and by modelling this biological interaction between the molecule and biologic targets, then we are able to bring a lot more data and can lead to better generalization.

Three steps of ComboNet model:

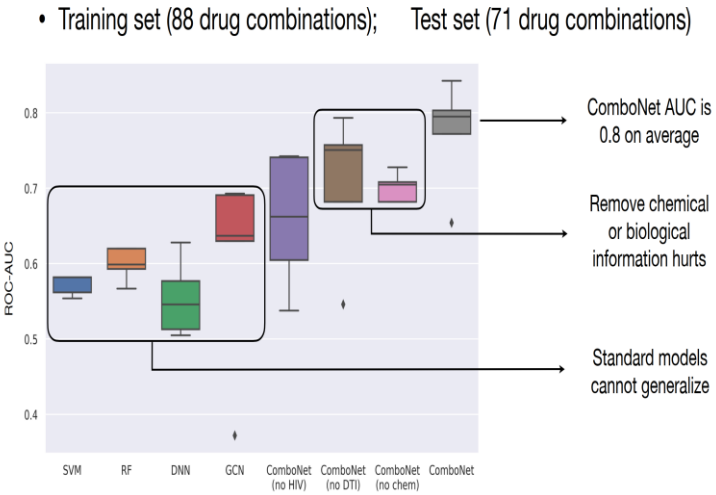
Figure source [7]

This model predicts the drug target interaction, like whether this compound inhibits certain biologic targets.



ComboNet model performance:

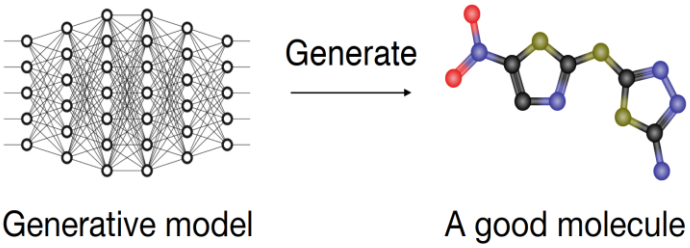
Figure source [7]



The performance of the model can be interpreted as, our best model on the right that includes both the biological knowledge and the chemical representation and if we just use standard models based on like SVM or deep neural networks without any of these biological knowledge, they just simply cannot generalize that well and also if we remove either the chemical or biologic information it will hurt the model quite a bit, so this can explain both biology and chemistry are important for the prediction path.

Graph Generation:

Figure source [7]

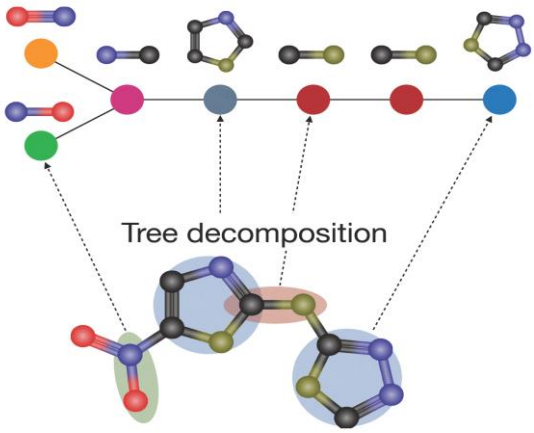


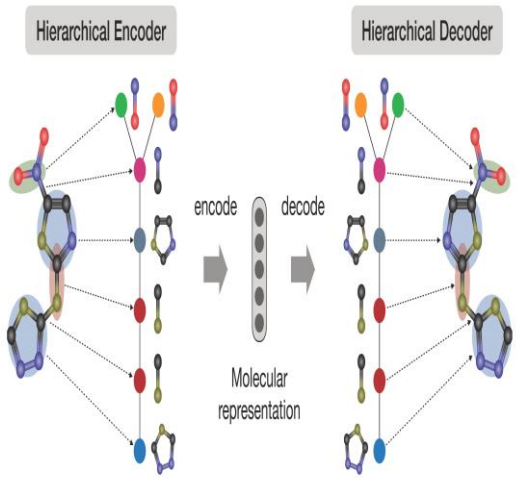
If we look at the previous approaches, we train GNNs, we rank all the compounds and we do experimental validation, it is good because it is fast, but it cannot scale up to the entire chemical space. We cannot do virtual screening if we want to explore the entire chemical space. So that's why we need totally different approach and that's where the graph generation comes in, as we want to learn distribution whose mass is concentrated around good molecules and if we can train these general models that can just directly sample good molecules from our model. So that we can efficiently explore the entire chemical space by doing this efficient sampling, but it is challenging problem, because we need to now generate graph structures rather than numerical numbers that standard machine learning models.

Junction Tree Variational auto-encoders: This model utilizes low tree waste prior of molecule, this method is inspired by the junction tree algorithm in the classic like graphic models and to leverage this low tree width of the molecule, we can decompose a molecule into a tree structure called junction tree.

What this junction tree does is, each node is no longer just a single atom, it is a motif or sub graph, because the molecules have no tree width each sub graph is relatively small, so tree width of a graph is basically the largest motif that we can get out of these three decomposition steps. Here the molecule has like tree with almost five because there's five-member ring and that's the largest and we can decompose rings further, but they are chemically meaningful, so this is what is known as low tree width. So, if we do this 3-D composition, we end up having these sub graphs that are small and another aspect is if we apply this treated composition over a large collection of molecules it doesn't result in that many types of sub graphs.

Given these two observations we can design a new type of variational auto-encoder that looks likes following.





So, we can map molecule into a continuous representation that incorporates this junction tree group that builds on this intermediate junction tree representation. Hence, we now we will embed both the molecule and the junction tree into this low dimensional vector and then we can decode them back to the molecule, but we will first decode the junction tree and then decode the molecule. This is advantageous because trees are much easier to generate.

Hierarchical graph encoder model architecture: Encoder is different from standard graph neural networks because we have both the graph and the junction tree, to be specific we can run graph convolution in this graph which learn future vectors for each atom, then we can propagate these atom vectors into the junction tree and run graph convolution there. So now each sub graph receives sub graph vector that represents how these sub graphs are connected together.

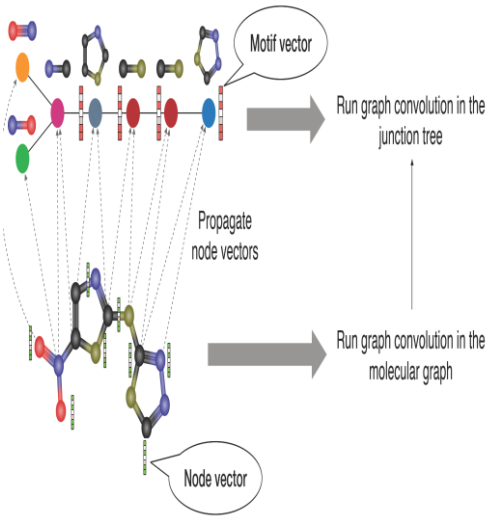
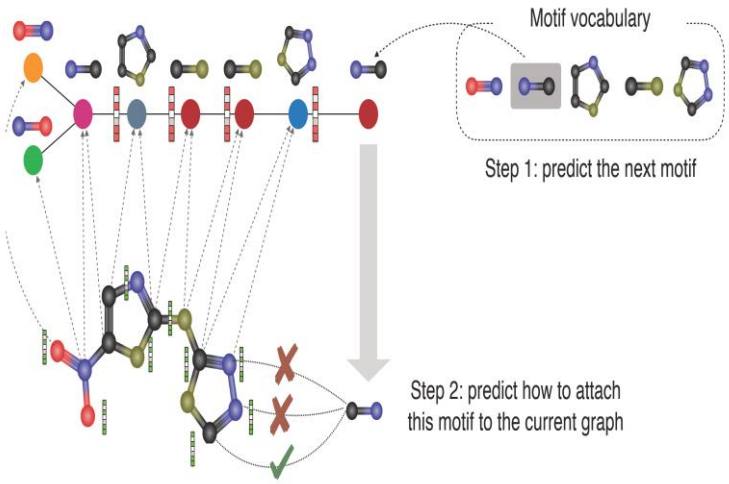


Figure source [7]

Now based on these encoded representations now we can generate a graph in a following way, in each step the model decides which sub graph to add next, so basically, we need to pick what's the next sub graph(motif) then we need to add to the current molecule that we are expanding, this is simply a classification task, so we can pick a single bound into the molecule.

Figure source [7]



Then we can predict how to attach this sub graph to the current graph by predicting what's the right attaching points between this sub graph and the existing graph.

Therefore, in this case we should attach atom (right corner ring atoms)

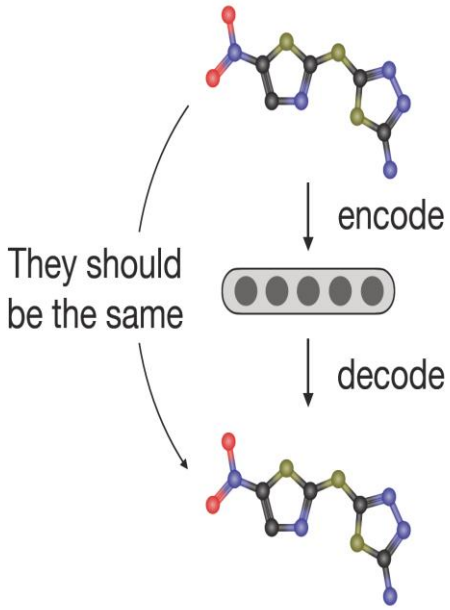


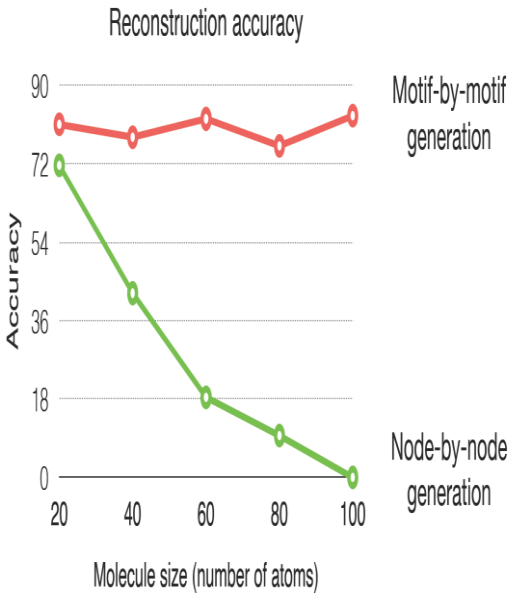
Figure source [7]

Then we can attach sub graph to the graph as above figure and this gives us one step procedure which is one step expansion of the molecule, and we can repeat this process multiple times until the model decides to stop and reason why this approach is advantageous because now, we are generating a sub graph by sub graph.

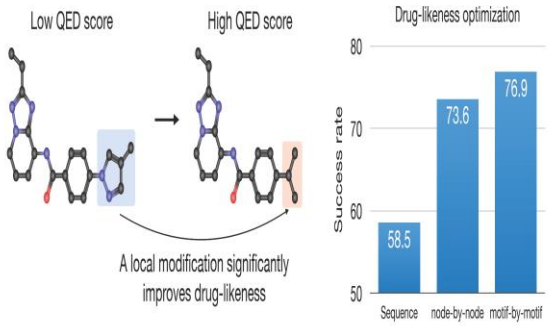
The time complexity is actually linear because we are generating this tree instead of generating this graph. Hence generating a tree, we can do this in a linear number of steps.

Experiment results:

The red curve represents the reconstruction accuracy of this sub graph by sub graph generation model, and we can see that the reconstruction accuracy remains high, if we have a large molecule or even if we have molecule which is up to 100 atoms, this is not common for drug-like molecules, but this demonstrates the advantage of the sub-graph by sub- graph generation.



Results on molecular optimization: This relates to the fine-tuning step. If we had some drugs that have like certain good properties, for example helison, but it may not be like drug-like enough, for example drug-likeness encompass many physiochemical properties and it greatly affects how it can be absorbed in the human body. So, let's say that helison is not good enough in terms of stroke likeness, then we can learn to modify this compound, by doing some local modifications so that it can improve the drug- likeness and therefore they can be better absorbed by human. So, this is where these generating models are good at, this way we can not



only generate molecules from scratch, but the model can also learn to modify existing graphs of this existing molecules to improve its physiochemical properties, while keeping it like antibacterial properties.

In this simple experiment drug-likeness is measured by QED and we can just directly compute it, so in a way that every molecule that we generate, we can actually compute its QED score quite easily. Therefore, as we can see above figure, this sequence space method does bad around 58% because of the wrong representation. Therefore, it shows the advantage of this multi-base generation

Refences:

[1] arXiv:1708.08227

[2] http://arxiv.org/abs/1510.02855v1

[3] *CS Cent. Sci.* 2018, 4, 1, 120–131Publication Date: December 28, 2017 https://doi.org/10.1021/acscentsci.7b00512

[4] Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. Science. 2018 Jul 27;361(6400):360-365. doi: 10.1126/science.aat2663. Epub 2018 Jul 26. PMID: 30049875.

[5] arXiv:1910.10685

[6] arXiv:1705.10843

[7] Wengong Jin, Deep learning for drug discovery, Massachusetts Institute of Technology