Specific aims:

Problem statement: The current state of biopharmaceutical research and development is a time-consuming process, costing billions of dollars and decades of research. This is due to the existence of tens of thousands of initial compounds, out of which very few are found to actually target the desired protein. Each potential drug must be tested, which drastically increases the price of the drug. In order to reduce the need for wet-lab testing, computational approaches can pre-screen drug candidates and accurately select promising ones to expedite the process. Although non-AI computational processes can also screen drug candidates, they are expensive, so they are not a worthwhile replacement for wet-lab testing. As Artificial Intelligence becomes stronger and more accessible, it may be the answer in reducing the effort needed to bring new drugs to the market.

Research question: Do artificial intelligence methods predict drug-protein binding affinities more accurately and take less time to train than non-AI, computer based methods? If so, which AI method and protein and compound representation will produce the most accurate binding affinity predictions and take least time to train?

Main goals of this research:

To test whether AI models can replace traditional non-AI methods in designing effective drug treatments: Binding affinity is the strength of interactions between a drug and protein. Predicting binding affinities between potential drug compounds and target disease proteins accurately is vital to designing effective drugs that can cure diseases [8]. If AI models can predict binding affinities more accurately or just as accurately as non-AI methods, then it could replace non-AI methods and reduce cost and time of drug discovery because AI methods do not require as much time or money learning known binding affinities.

To test which AI method has the most potential in drug discovery: Some AI methods may predict binding affinities more accurately than other methods. Comparing the accuracy of the binding affinities given by multiple AI types in our research gives us information about which AI method has the most potential for drug discovery.

Minor goals:

Test whether the protein and compound representation type impacts accuracy of binding affinity predictions: It is still unclear whether the computer representation type of proteins and compounds impacts AI and non-AI methods' accuracy of predicting binding affinities. Our experiment will touch on whether the way that proteins and compounds are computationally represented impacts each method's accuracies of predicting binding affinities.

Determine how much training time each AI model requires:

Pharmaceutical companies want to use computer models that not only can predict drug binding affinity accurately, but also be trained quickly. Our experiment will determine how much training time is necessary for each of the AI models we will be using.

Hypothesis: AI models will be able to predict drug binding affinities more accurately than non-AI computer methods. Specifically, the most accurate AI method will be a Graph Convolutional Network (GCN) paired with 2-D Convolutional Neural Network (CNN) representations of both proteins and compounds. Among the AI models, the training time for GCNs with 2-D protein and compound representations will also be the lowest.

Background:

Drug Discovery: The current drug discovery process takes an average of 10-15 years and can cost up to 2.6 billion dollars to develop one new prescription drug from thousands of compounds [1]. This is largely due to the number of initial early failures in the drug discovery process, as well as later issues of safety and efficacy. This imposes the need to improve the efficiency of drug discovery by pre-screening drug molecules and accurately selecting promising candidates for further testing and clinical trials. The drug discovery procedure typically begins by identifying the target protein implicated in diseases followed by determination of compounds that bind to the protein with maximal wanted effect and minimal side effect [2]. However, this binding is dependent on the complex structures of both molecules, and it is nearly impossible to predict if a ligand will bind, much less determine a binding site. Because of the lack of efficient tools to pre-screen drug molecules for promising candidates, most pharmaceutical companies still perform large amounts of laboratory testing.

Computer Aided Drug Discovery: Current Computer Aided Drug Discovery (CADD) methods include real-time analysis of properties between proteins and ligands in a solution, with each molecule being modeled by the computer [3]. However, this requires a previous understanding of the protein and ligand, specifically the attractions between atoms, which must be inputted beforehand. This process is also computationally intensive, requiring supercomputers to simulate even nanoseconds of real time [4].

Artificial Intelligence may solve the issue of needing to pre-input relationships between molecules, as it may be able to learn these underlying properties by itself. There are currently no Artificial Intelligence models that are able to accurately predict protein-ligand binding to an extent at which they may be usable for pharmaceutical applications, but there are starting points from which we can begin our search. We will be using a variety of AI models proposed by several researchers that are proven to be able to predict drug binding affinities, such as Convolutional Neural Networks (CNNs) [3] [7], General Adversarial Networks (GANs) [5], and Graph GCNs (also known as spatial transformers) [6, 7]. Our research

expands upon past research by comparing the prediction performance of all these AI models with non-AI methods. It also takes into account the impact of computer protein-compound representation type on the accuracy of drug binding affinities and training time, which is mostly not covered in previous research. Our research aims to create a more robust and refined conclusion about which combination of computational method and protein-compound representation results in the highest-accuracy predictions and lowest training time.

Research Design and Methodology:

Experimental Design:

Experimental goal: Our aim is to determine the efficacy of the semi-supervised GANsDTA, supervised 1-D CNN, and GCN model's predictions for drug binding affinity compared to accuracy of computer-aided ligand-based methods (CADD) for various computer protein and compound representations. The training time of each type of model will also be observed. We expect that AI models, especially GCN with 2-D protein and compound representations, will predict binding affinity more accurately and take less time to train than non-AI models.

Independent variables: type of AI model (CNN, GANs, GCN), protein and compound representation type

Dependent variables: Training time measured in days and models' prediction accuracy of drug binding affinity measured by a scoring function/power from 8-18. The scoring function is a predicted value of drug binding affinity and the MSE is how much the scoring function can vary from its average.

Controlled variables: The dataset of proteins and compounds that each model will use to predict protein-ligand affinity.

Control group: non-AI Computer Aided Drug Discovery methods (ligand-based interactions) [4].

Materials: AI models (CNN deep learning model, GCN model, GANsDTA model), non-AI control computer ligand-based methods(CADD), computer, KIBA and Davis datasets listed in Experimental Procedure section below

AI model details and training instructions:

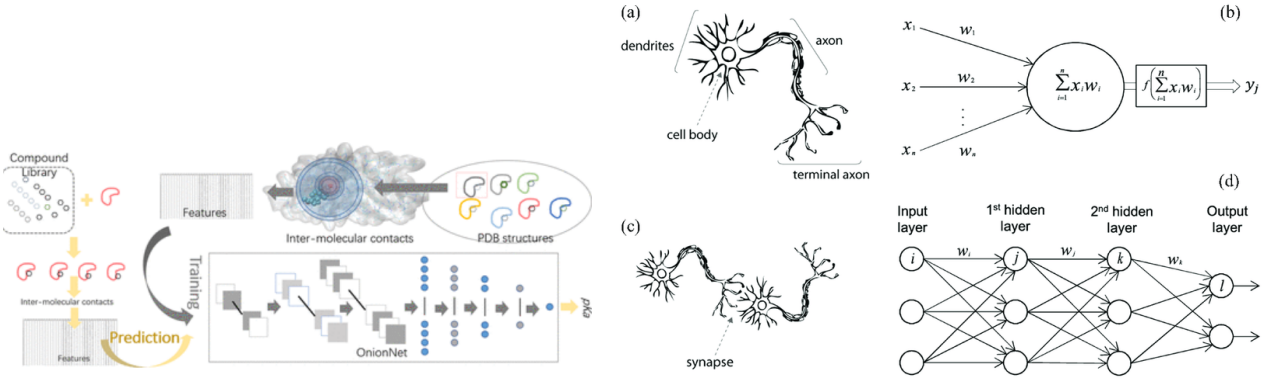
For all models: Train each model to computationally represent proteins and compounds in the ways specified in the Predictions step of the Experimental Procedure section below. Assume that each model can learn 65-150 protein-compound pairs at a time. Record the time needed to train each model in the tables below in days. Each model can be built and trained on computers, so training is realistic to do anywhere and does not require in-person interaction, making this experiment feasible regardless of the state of COVID-19.

Convolutional Neural Networks: CNN deep learning models [3, 5] are supervised and can be either 3-D or 1-D.

Supervised 1-D neural networks are trained to predict drug compound and protein affinity based on **pre-classified** protein

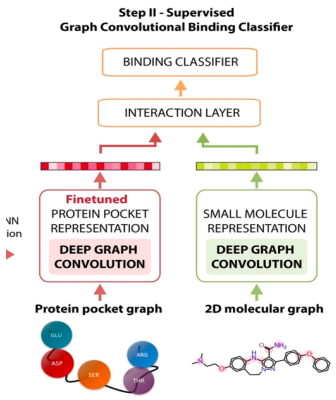
DNA sequences and simplified, linear molecular structures of compounds. For this experiment, only the 1-D CNN method will be used because some 3-D structures of protein-ligand structure are unknown, too much memory is required in 3-D neural networks, and parameters become overly complicated in multidimensional models. In a 1-D model, weights of artificial neurons in neuronal networks (w1, w2, w3 Figure 1 right) are adjusted, so that their sum (sigma notation in rectangle Figure 1 right) is strong enough to overcome the threshold (circle) and fire action potentials, thus modeling learning in biological neurons. A supervised 1-D model is trained so that two CNN blocks initially detect patterns in protein DNA sequences and compounds that are already classified. Subsequent layers of connected neural networks further analyze and combine data from CNN blocks and layers below in a hierarchical manner to form complex representations of proteins and ligands (compounds). The CNN can predict drug binding affinity from examining protein and ligand interactions. A problem is that this supervised neural network cannot classify unlabeled data like unsupervised GANs can, so it will cost a significant amount of money and time to pre-classify protein and compound sequences before inputting them into CNN.

Figure 1: CNN model example OnionNet [3] and Artificial Neuron Learning



Graph Convolutional Networks: GCNs (also called spatial transformers) [6, 7] are a unique subclass of **supervised CNNs** that can possibly predict drug binding affinities better than 1-D CNNs because they are able to learn from more complex 2-D molecular structure and symmetry. Each neuron is connected to other neurons in the previous layer through a set of weights that are adjusted to simulate action potentials to overcome threshold and induce learning (Figure 1 right) . The first **interaction neuronal network** layer (Figure 2) is trained to directly detect and extract simple spatial features from 2-D molecular data about proteins and compounds by modeling atoms as nodes and bonds as angles. The remaining neuronal layers or **binding classifiers** (Figure 2) are structured in a hierarchical manner similar to 1-D CNN models, so that each subsequent layer can be trained to combine simple 2-D features from previous layers into more complex features and larger substructures of the protein-compound complexes. The GCN is able to make better predictions when there are complete spatial representations of the protein and ligand.

Figure 2: Graph Convolutions and 2-D Representations of Proteins and Compounds [6, 7]



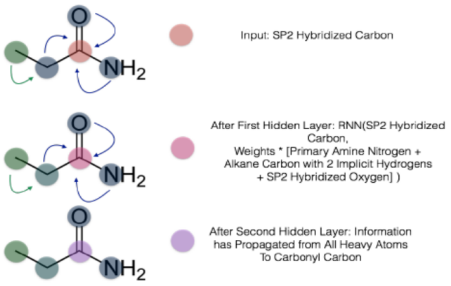
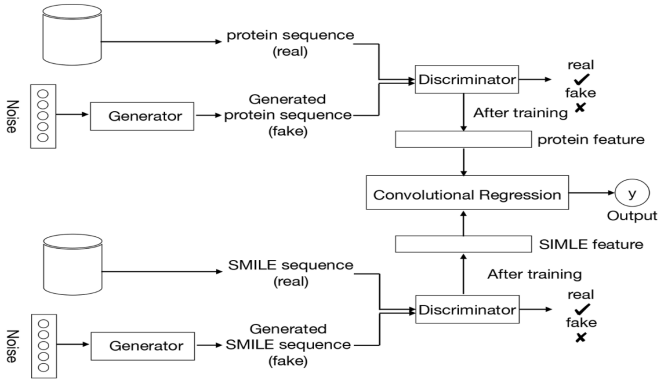


FIG. 1. Visual depiction of Gated Graph Neural Network with atoms as nodes and bonds as edges. The small molecule propanamide is chosen to illustrate the propagation of information among the different update layers of the network.

General Adversarial Networks: Semi-Supervised GANsDTA [5] (Figure 3) are trained to have an unsupervised GANs system and supervised CNN. Unsupervised GANs have two differentiable functions: a generator and discriminator. A **generator** is trained to learn in an unsupervised manner from unclassified data to produce fake protein DNA sequences and drug chemical compounds. Similarly, the **discriminator** is **unsupervised** in that it is able to distinguish if the protein sequences and compounds come from the fake **generator** or the actual dataset without predefined categories of which data is fake or real. Using unsupervised methods like GANs can potentially eliminate the cost and effort needed to make labeled data of protein and drug compounds, which is necessary in supervised methods. However, **supervised 1-D regression CNN** is necessary to make predictions of drug binding affinity based on actual protein sequences and compound representations that are classified by the discriminator. This CNN will be trained just like an independent 1-D classification CNN as described above, but the difference is that independent 1-D CNN require the expensive and labor-intensive process of pre classifying data before inputting them into the model (supervised learning), whereas a joint GANs-regression 1-D CNN can make predictions about binding affinity based on protein DNA sequences and compounds already classified by unsupervised GANs learning methods, eliminating the need for supervised learning.

Figure 3: GANsDTA [5]



Control baseline non-AI method (CADD): Computer ligand-based interactions [3, 4, 6] will be trained to use data about known binding relationships of compound-protein pairs from **Pub-Chem website** [9] to predict binding affinity between similar pairs in the **KIBA and DAVIS datasets below** [5]. This however requires supercomputers to simulate even nanoseconds of real time, and thus our baseline may be less accurate than the most up to date methods.

Experimental procedure:

1. Data collection of protein-compound representations:

Each model will computationally represent proteins and compounds as different visual forms based on the benchmark datasets obtained from an earlier research study about GANsDTA [5]. This study's proteins and compounds are stored on an open, free chemistry database called **Pub-Chem** [9]. 1-D CNNs can either visually represent proteins as protein DNA sequences using the **Smith Waterman algorithm** [5], which selects and compares similar regions of DNA sequences rather than the entire sequence, or DNA sequences in 1-D CNNs form (see training instructions). 1-D CNNs can represent compounds as more complex molecular structures found on Pub-Chem databases or simplified, molecular 1-D CNN forms. Similarly, GCNs can represent compounds as either molecular structures stored in the Pub-Chem database or build 2-D compound representations from 2-D CNNs (see training instructions above), but it has to represent proteins as 2-D CNN forms. Computer-based ligand methods (CADD) can only use the Smith Waterman algorithm to represent known protein DNA sequences and the Pub-Chem website [9] to represent compounds**.** Assume each model can analyze 65-150 protein-compound pairs at a time. Not every protein-compound pair has to be tested given the time-consuming process, so testing around half of the KIBA compounds is acceptable.

	Proteins	**Compounds**	**Interactions**
KIBA (Pub-Chem)	**229**	**2111**	**118254**
DAVIS (Pub-Chem)	**442**	**68**	**30056**

2. Model/Parameter Training:

First, train each Al model (1-D CNN, semi supervised GANsDTA, graph convolutions (GCN)) and the control ligand-based model according to the training instructions above. Train each model to represent proteins and compounds in the ways listed in the Prediction step's tables below until the model is able to predict binding affinity. Record the training time (in days) in the tables. Observe how the time differs according to the type of training model. The fact that training each model takes days, weeks or even a month reflects the reality of most AI-based research to create functional AI models that

can make accurate predictions. Training each model separately at the same time reduces the overall time needed to conduct the experiment. Adam's optimizer is an algorithm automatically encoded in each model that sets the initial learning rate of each model to 0.001 [5]. However, it may be necessary to adjust the learning rate or other parameters such as the depth of the neural network or the momentum of the network.

3. Prediction:

After training each method and recording, we will then conduct separate comparison tests on each of the KIBA and DAVIS datasets. For each test, have each model represent proteins and compounds in the ways listed in the tables below (e.g. 1-D CNN method with a Smith Waterman protein representation and Pub-Chem compound representation). Refer back to the data collection section for what Smith Waterman or Pub-Chem is. When each model is done studying the protein and compound using different representations, the model will automatically predict the **average scoring function value** (range from 8 to 18) and **MSE**(just a measure of how much the actual scoring function can vary from the average value) for each group of 70-150 protein-compound pairs in each database. Average the values of the all groups to represent the average of each dataset (may not represent every but most protein-compound pairs). The algorithms for the scoring function and MSE are already encoded into each model method, so there is no need to manually input them. Record these below.

Set up two tables: one based on KIBA dataset and other based on DAVIS dataset

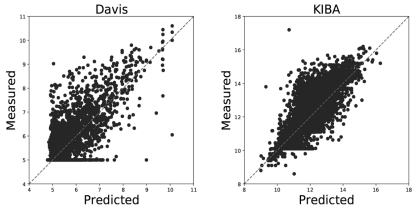
Method	Protein representation	Compound representation	Average Predicted Scoring function (8-18)	Average Actual scoring function (8-18)	MSE	Experimental error	Time needed to train model (days)
1-D CNN	1-D CNN	1-D CNN					
1-D CNN	Smith Waterman	Pub-Chem					
1-D CNN	Smith Waterman	1-D CNN					
GCN	2-D CNN	2-D CNN					

GCN	2-D CNN	Pub-Chem					
Semisuper-vised GANs	GANs	GANs					
Computer-based ligand methods (control)	Smith Waterman	Pub-Chem					

4. Calculations:

Calculate and record the experimental error between actual and predicted average binding affinity values (scoring functions in table above) of each model. The smaller the error, the more accurate the prediction.

Figure 4 [5] Predicted vs Actual Average Binding Affinities for KIBA and DAVIS dataset



$$\frac{|\text{Predicted binding affinity} - \text{actual binding affinity}|}{\text{actual binding affinity}}$$

Strengths and weaknesses of proposed procedure:

A potential weakness of our procedure is that our control non-AI methods did not include other methods besides computer based ligand systems because non-AI methods are computationally expensive and take more effort to train. If we could test the accuracy of other types of non-AI methods, then we would have a better idea of whether Artificial Intelligence really predicts drug affinities more accurately. Another weakness is that we could not compare actual and predicted binding affinities of each individual compound-protein pair since that would be too time consuming with larger datasets. Larger datasets are necessary because they provide more accurate mean binding affinity values and smaller margin of errors. However, the good thing is that our procedure takes into account multiple variables, such as the accuracy of predicted binding affinities and training time for each type of model and protein-compound representation. This gives pharmaceutical companies a thorough overview that they can use to decide whether AI models can replace computer-based systems, which AI model to use, and how proteins and compound yields can be represented in a model. Moreover, we performed the experiment procedure on two datasets instead of one to account for errors in measuring a model's scoring functions for one

dataset. We also ensured that the experiment would be realistic since our models can be easily trained on computers. Datasets of proteins and compounds can be found on the free Pub-Chem database.

Data Analysis: If the predicted average predicted binding affinity value of AI models is closer to actual average binding affinity value (measured by the equation in Figure 4) than the prediction of computer-based ligand methods, then our data supports the main part of our hypothesis that AI models predict binding affinity more accurately than non-AI methods. This is because ligand-based methods require data of known drug and protein binding affinities to predict binding affinities of similar drug-protein pairs, so it is not possible to predict binding affinities accurately when interactions between a drug and protein are different. By contrast, AI models are able to predict binding affinities without known binding affinity values by hierarchically building complex representations of proteins and compounds from simple data features. Thus, AI models could potentially replace computer models in designing effective drugs to target disease protein. On top of being more accurate, the shorter training times of AI models suggest that they do not have to spend time and money learning data of known binding affinities. AI's potential to discover drugs that can cure disease is a significant advancement from traditional AI purposes in cognitive science, such as neural language processing and computer vision. Among the AI models, if the experimental error is least for graph GCNs with 2-D protein and compound representations, then data also supports the second part of the hypothesis that some AI methods yield more accurate predictions than others. This is because graph GCNs are able to see protein-compound interactions more clearly from complete 2-D molecular protein and compound representations than 1-D or Smith Waterman protein DNA sequences and simplified molecular compounds in 1-D and GANsDTA models. The training time is also least for graph GCNs because 1-D CNNs and GANsDTA models take more time to analyze interactions between linear protein DNA sequences and simplified compound structures since it cannot directly visualize 2-D or 3-D representations, whereas graph GCNs can more directly understand the spatial aspects of protein and compound from their 2-D molecular structures. Pub-Chem compound representations are not as good compared to 2-D compound representations because it not only takes more time to train models to represent compounds similar to ones on Pub-Chem, but Pub-Chem compound representation may be inconsistent with 2-D protein representations. Training time should also be taken into consideration when determining which model is the best possible model for pharmaceutical companies to use. Therefore, an ideal model for predicting binding affinities would be graph GCNs and 2-D CNN representations of both proteins and compounds.

Ethical Implications: Because this study does not involve human subjects, there is no risk of infringing on human safety while conducting the experiment. However, the results of the experiment absolutely have ethical implications that must be considered. The goal of the study is to hasten and cheapen the drug discovery process, which will directly affect the

population of people with physical or mental illnesses that could be eased or cured by pharmaceutical drugs. It is important to ensure, therefore, that the results of our experiment do not put this vulnerable population at risk. For example, should AI prove to discover drugs with high accuracy, there may be incentive for pharmaceutical companies to use AI to cut costs in other ways, like replacing lab testing and current computer models. The FDA currently requires thorough lab, animal, and human testing of any drug before it is approved and allowed on the market, so this risk is not one we currently have to consider, but it may be in the future when AI becomes more advanced. Additionally, our goal with this experiment is to cut costs for the consumer, but the pharmaceutical industry may not share that goal. If drugs cost less to discover using AI, the pharmaceutical industry could use that to increase profit margins for themselves rather than decrease cost for the consumer. This will not technically put consumers through increased financial harm, but it will not be helping them as the study intended. At the very least, using AI in the drug discovery process will cause pharmaceutical companies to lose their biggest argument for high markups, which is that extra profit is needed to fund the discovery of new drugs [10]. Without that argument, price regulation legislation stands a better chance of being passed, allowing the results of our experiment to help consumers as intended [1].

Limitations: As mentioned in the Strengths and Weaknesses section of the procedure, the control group consists of only one non-AI method. Therefore, the success or failure of each AI model is only determined in the context of that particular method, which does not allow us to extrapolate to conclude that AI is more or less successful than non-AI methods as a whole without further experimentation. Additionally, each of the three AI models in our experiment either is supervised or has a supervised component. The process of labeling data to train the supervised components is time-intensive and must be done by humans. So, if the types of AI in this study turn out to be more successful than the non-AI baseline and eventually are used in the drug discovery process, it would not entirely eliminate the need for humans in the drug discovery stage, nor would it cut down completely on the time needed for the process.

Future implications: Our research demonstrates that AI's ability to simulate cognition and learning gives it an advantage over non-AI methods when it comes to helping humans advance the medical field. In the future, we hope to refine AI models so they play a bigger role in organic chemistry, such as being able to model and analyze synthesis and decomposition of chemical compounds, atoms and bonding in chemical structures, and dimensional analysis calculations [8].

Works Cited

[1] DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of health economics, 47, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

[2] Settleman, J., & Cohen, R. L. (2016). Communication in Drug Development: "Translating" Scientific Discovery. Cell, 164(6), 1101–1104. https://doi.org/10.1016/j.cell.2016.02.050

[3] Zheng, L., Fan, J., & Mu, Y. (2019, September 16). *OnionNet: a Multiple-Layer Intermolecular-Contact-Based Convolutional Neural Network for Protein–Ligand Binding Affinity Prediction*. ACS Publications. https://pubs.acs.org/doi/10.1021/acsomega.9b01997.

[4] Leelananda, S. P., & Lindert, S. (2016). Computational methods in drug discovery. Beilstein journal of organic chemistry, 12, 2694–2718. https://doi.org/10.3762/bjoc.12.267

[5] Zhao, L., Wang, J., Pang, L., Liu, Y., & Zhang, J. (2020, January 9). *GANsDTA: Predicting Drug-Target Binding Affinity Using GANs*. Frontiers. https://www.frontiersin.org/articles/10.3389/fgene.2019.01243/full.

[6] Feinberg, E. N., Sur, D., Husic, B. E., Mai, D., Li, Y., Yang, J., … Pande, V. S. (2018, March). *Spatial Graph Convolutions for Drug Discovery*. ResearchGate. https://www.researchgate.net/publication/323746994_Spatial_Graph_Convolutions_for_Drug_Discovery.

[7] Torng, W., & Altman, R. B. (2019, October 3). *Graph Convolutional Neural Networks for Predicting Drug-Target Interactions*. ACS Publications. https://pubs.acs.org/doi/10.1021/acs.jcim.9b00628.

[8] de Almeida, A. F., Moreira, R., & Rodrigues, T. (2019, August 21). *Synthetic organic chemistry driven by artificial intelligence*. Nature Reviews Chemistry. www-nature-com.libproxy.berkeley.edu/articles/s41570-019-0124-0.

[9] Kim, S., Chen, J., ...Bolton, E. E..(2019). PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research*, *49*(D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971

[10] Emanuel, E. J. (2019, June 26). *Big Pharma's Go-To Defense of Soaring Drug Prices Doesn't Add Up*. The Atlantic. https://www.theatlantic.com/health/archive/2019/03/drug-prices-high-cost-research-and-development/58525