Project Code And Title

Project Title: The design & construction of an iterative process using two machine learning models to generate the best inhibitors for a target protein

Project Code: F41

Student Name: Tanush Goel

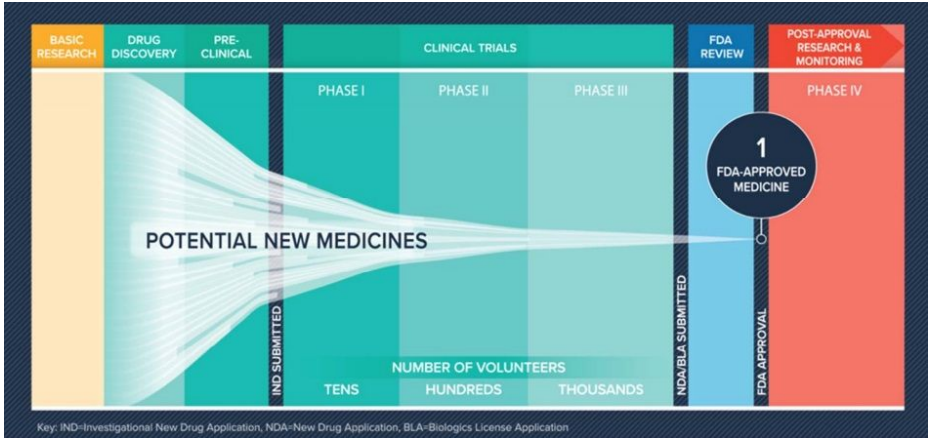
RRI Project: No

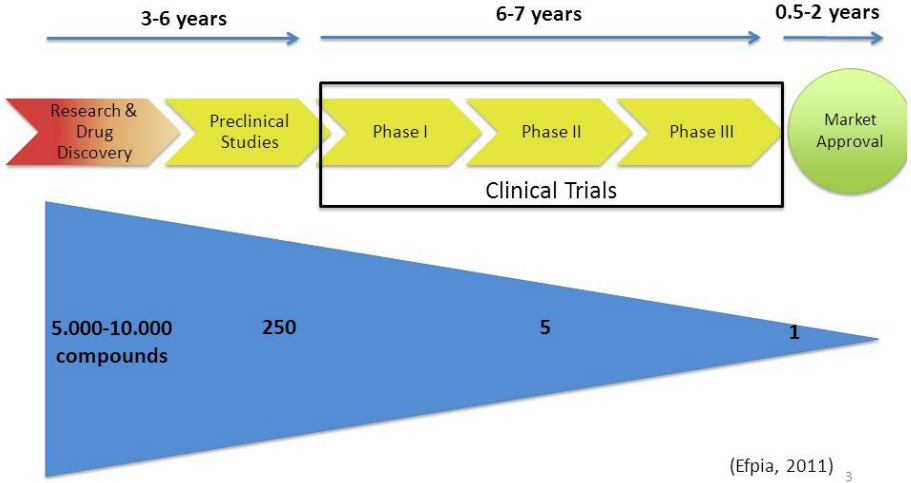
Grade: 11

Introduction

Problem: Drug Discovery is a VERY long and expensive process

- Average time from FDA application to drug approval ~12 years
- Average cost of taking a drug from concept to market exceeds $1 billion
- Of up to 10,000 compounds tested, only 1 will reach the market



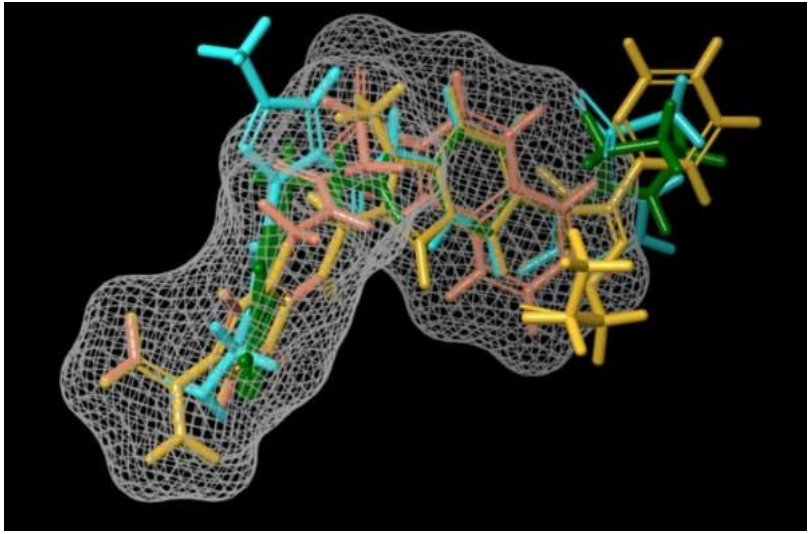


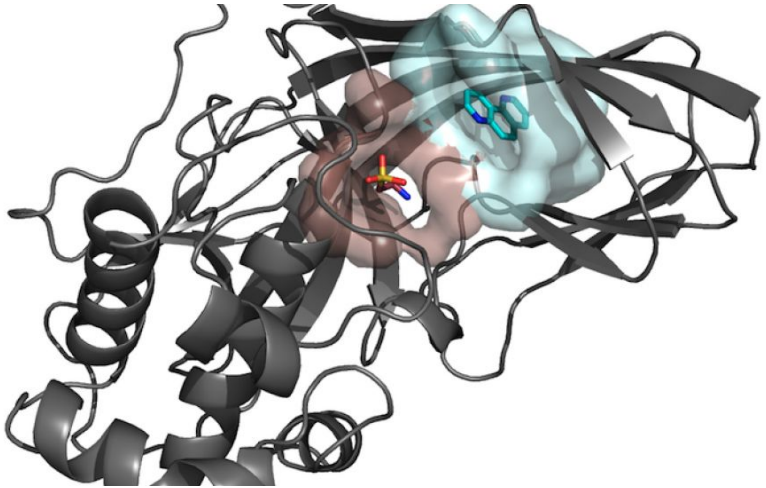
(Efpia, 2011)

Introduction

Existing work to solve this problem:

- Physics-based computer simulations
  - Pros:
    - Lets scientists model drug-target interactions
  - Cons:
    - The protein structure needs to be solved before simulation can take place
      - Near useless for new viruses
    - Computationally expensive, taking days or even weeks to perform one simulation
    - Can be inaccurate or lead to no results





Introduction

Engineering Goal:

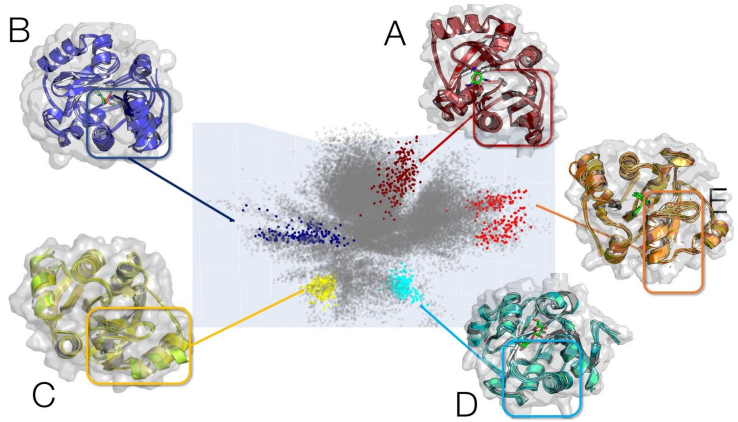
Use artificial intelligence to create the best inhibitors (drugs) for any given target protein to help reduce the time and cost of the drug discovery process

Currently, AI fails to outperform pharmaceutical scientists in creating viable drug candidates

A report by the HIMSS Analytics Essentials Brief shows that less than 5% of healthcare organizations are currently using or investing in AI technologies

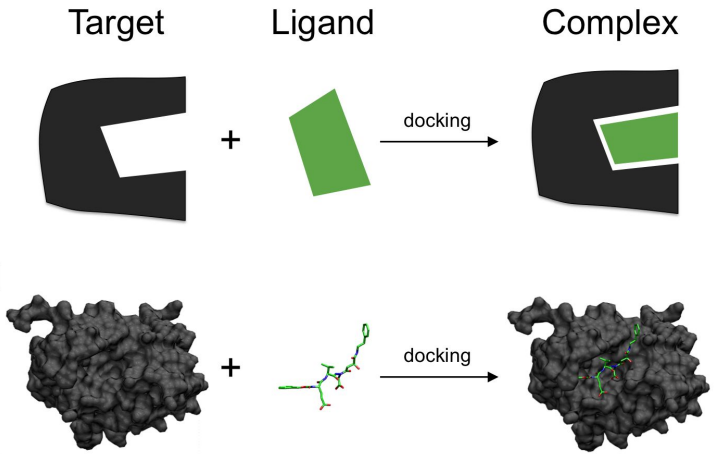
Methods

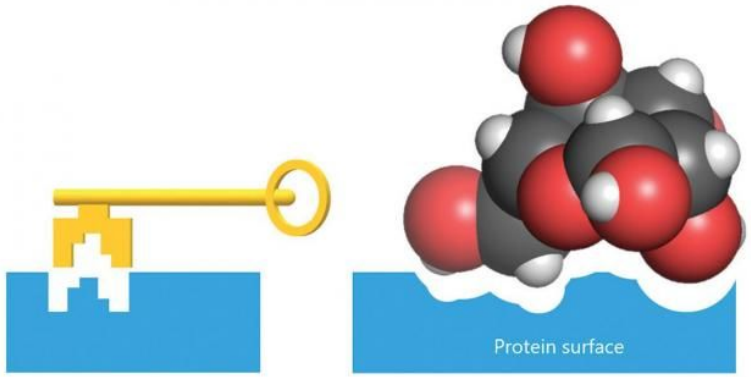
- 3 Major Steps:
  1. Create a neural network to generate new drug-like compounds
  2. Create another neural network to predict how well these compounds will bind to a certain protein
  3. Create an iterative process to generate the best drug-candidates
     a. Generate/augment compounds with the first model
     b. Score the compounds based on how well they bind to the target protein with the second model
     c. Take the best 50 compounds
     d. Start a new iteration



Methods

- Drug-Target Interactions:
  - When a virus infects host, it uses host cells to make its own viral proteins
  - Drugs are made to inhibit these proteins (prevent them from functioning)
    - To do this, the drug/ligand must bind to the target protein very well
  - Like finding a key for a lock, we are trying to find the drug that best fits the protein surface

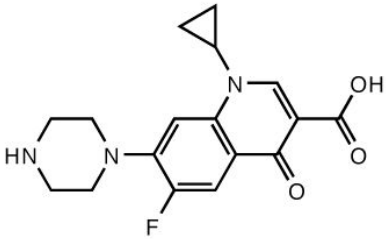




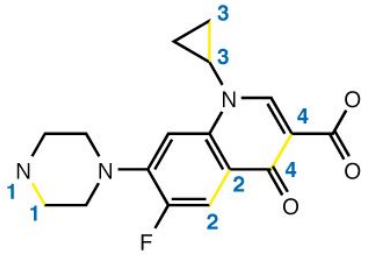
Methods

- SMILES (Simplified Molecular-Input Line-Entry System)
  - Atoms are shown by atomic symbols
  - Hydrogen atoms are assumed to fill spare valencies
  - Adjacent atoms are connected by single bonds
    - double bonds are shown by "="
    - triple bonds are shown by "#"
  - Branching is indicated by parenthesis
  - Ring closures are shown by pairs of matching digits
- Represents 3D molecule as 1D string
- Example:
  - Penicillin can be represented as:
  - CC1(C(N2C(S1)C(C2=O)NC(=O)CC3=CC=CC=C3)C(=O)O)C

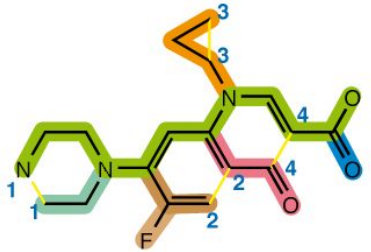
A



B



C

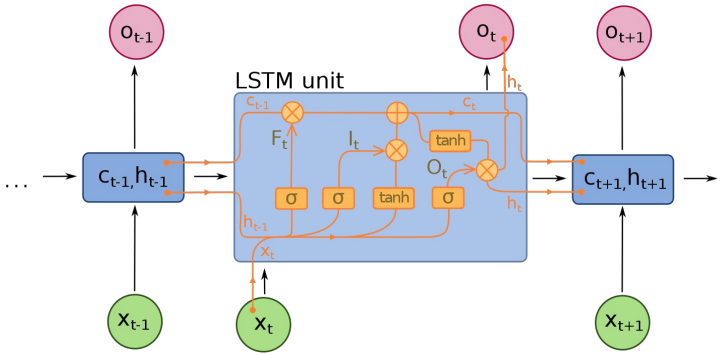


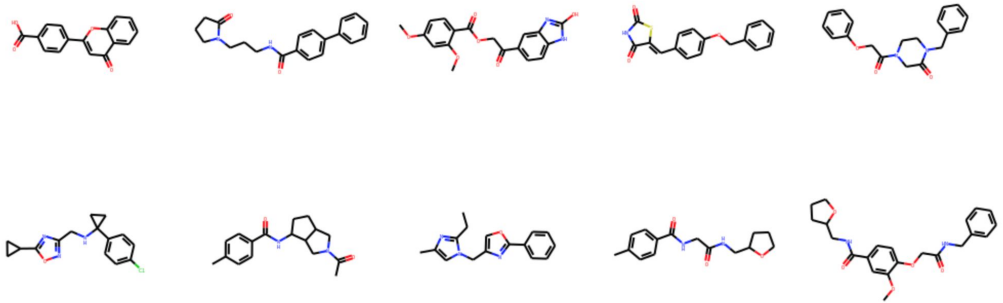
D

N1CCN(CC1)C(C(F)=C2)=CC(=C2C4=O)N(C3CC3)C=C4C(=O)O

Methods

- Model #1 (CuDNNLSTM neural network)
  - Uses LSTM (Long-Short Term Memory) cells to look at the previous 100 elements of a SMILES sequence and predict what element should be next
  - ...CC1(C(N2C(S1)C(S1)C(C what's next?
  - End token: "$" - tells us when the model thinks the compound is complete
  - "CC1(C(N2C(S1)C(C2=O)NC(=O)CC3=CC=CC=C3)C(=O)O)C$"
  - Doing this repeatedly allows us to generate novel compounds

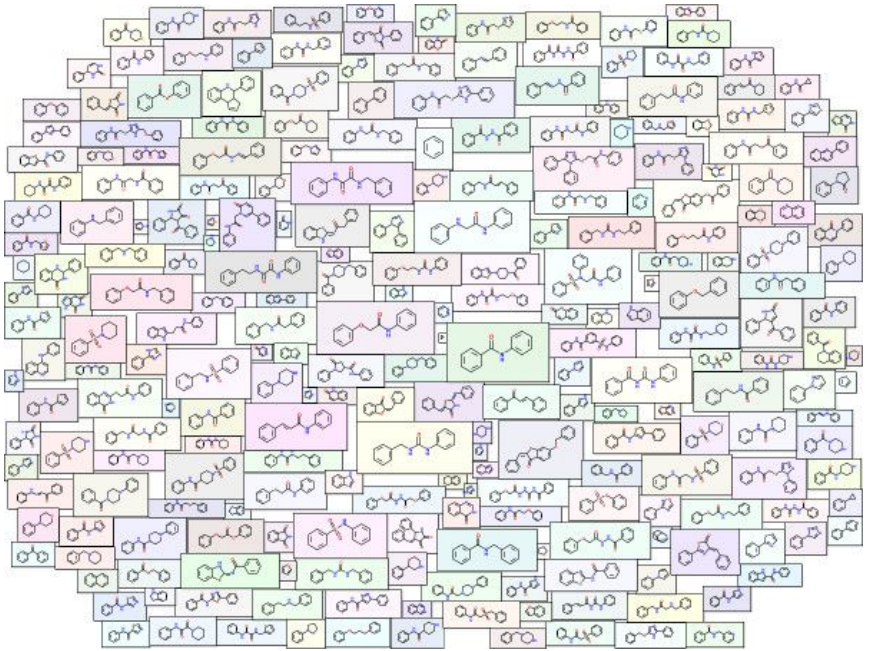




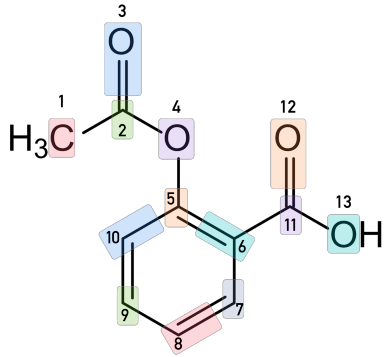
Methods

- Model #1 (CuDNNLSTM neural network)
  - Trained on >1.5 million drug-like compounds and validated on ~50k compounds
  - Collected data from ChEMBL
  - Training Accuracy: ~90.1%
  - Validation Accuracy: ~89.2%

ChEMBL



Methods



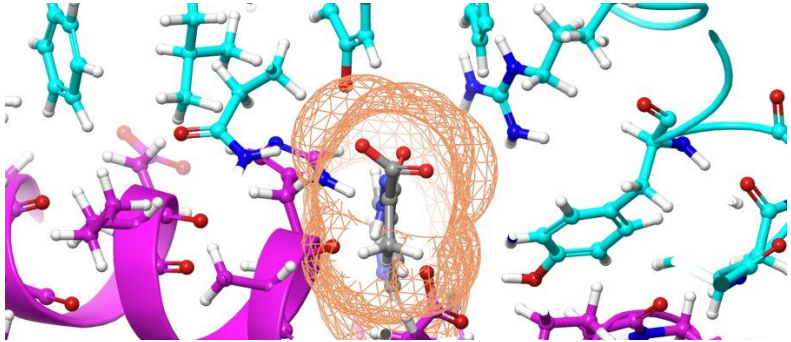
1 2 3 4 5 6 7 8 9 10 11 12 13
CC(=O)OC1=CC=CC=C1C(=O)O

Before outputting any molecule the model generates, it makes sure that:

1. The molecule is valid
   a. It follows the laws of chemistry and SMILES format
2. The molecule is drug-like
   a. It passes Lipinski's Rule of Five (a set of constraints in order to maintain drug-like character within the compounds - otherwise the drug may have poor permeation/absorption, faster rate of metabolism and excretion, unfavorable distribution, and might be toxic)
      - H bond donors (OH and NH) <= 5
      - Molecular weight <= 500 Daltons or g/mol (same thing)
      - Log P <= 5
      - H bond acceptors (N and O) <= 10
3. The molecule is unique
   a. It is not in the list of compounds it has already generated

MW < 500 Da
CLogP < 5
H-bond donor < 5
H-bond acceptor < 10

Methods



3 generation functions:

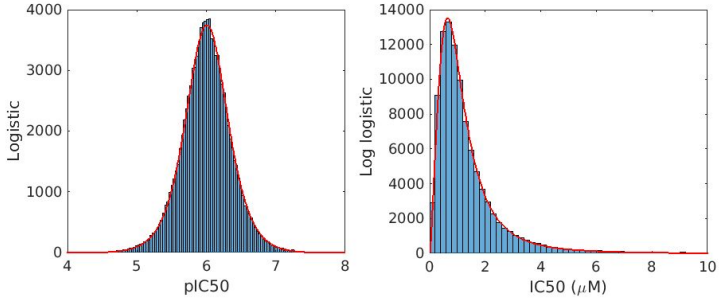
1. Generate Randomly
    a. Takes a randomly sampled sequence and builds off of it, outputting a random drug-like compound
2. Generate Methodically
    a. Takes a list of molecules, creates sub-sequences from them, and builds a molecule using each sub-sequence
3. Augment
    a. This technique replaces random parts of molecules with model predictions
        i. Makes sure the molecules are at least 20% similar

Generates valid, drug-like molecules ~82% of the time

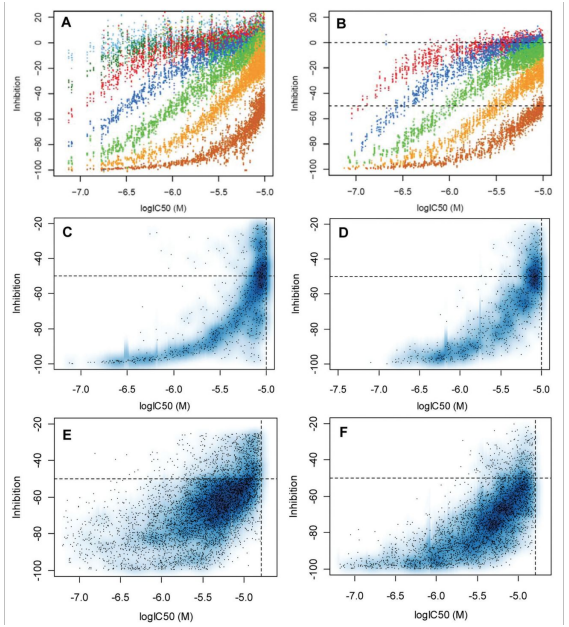
Note:

- Model predictions are used based on output probabilities
    - Example:
        - The model predictions for the next element are 50% for carbon, 20% for boron to be the next element, and 30% for sulfur to be the next element of the SMILES sequence
        - When generating a molecule, the next element will be chosen based on these probabilities:
            - Carbon has a 50% chance to be chosen as the next element
            - Boron has a 20% chance to be chosen
            - Sulfur has a 30% chance to be chosen

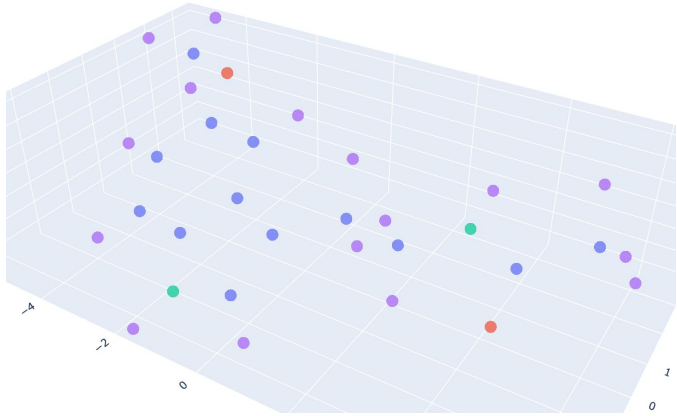
Methods



- pIC50
  - Tells us how well a drug will bind to a certain protein
  - IC50 (half-maximal inhibitory concentration) definition:
    - A measure of the potency of a substance in inhibiting a specific biological/biochemical function
    - Asks “how much of this drug is needed to inhibit this protein?”
      - The less of the drug needed, the better it must bind to the protein
  - pIC50 is the negative log of IC50 in molar
  - The best drug candidates are the ones with the highest pIC50



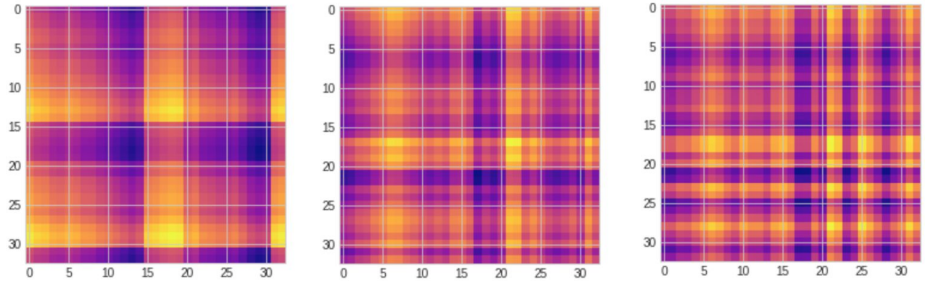
3D Coordinates For Melatonin:



Methods

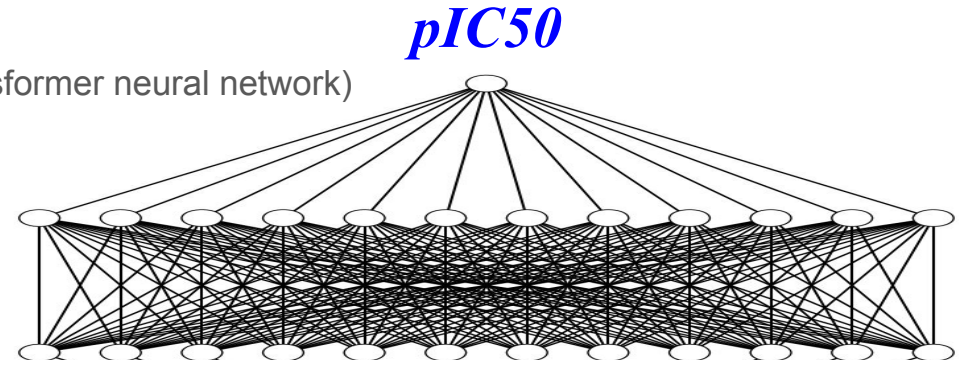
- Model #2 (CNN/Transformer neural network)
  - Predicts pIC50 of drug-target interaction
  - 3 inputs to the model:
    - Protein sequence in FASTA - uses 1D convolutions
      - FASTA (FAST-All) - amino acid sequence
      - “MEGTPAANWSVELDLGSGVPPGEEGNRTAGPPQRN…”
        - each letter corresponds to an amino acid
    - Drug compound in SMILES - uses transformer encoder blocks
    - Distance matrix created from the drug compound - uses 2D convolutions
      - Calculates distance between every x, y, and z coordinate for every atom of the molecule

Distance Matrix For Melatonin:

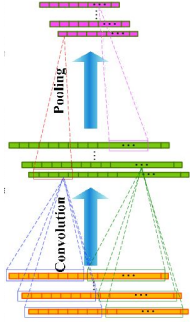


Methods

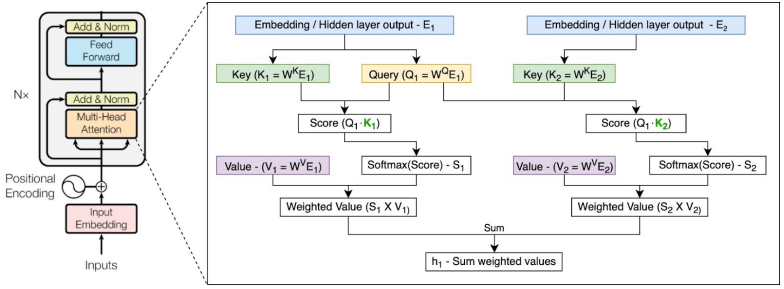
- Model #2 (CNN/Transformer neural network)



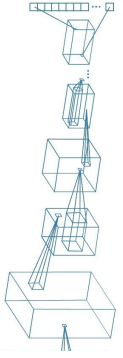
1D CNN FASTA

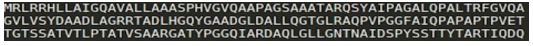


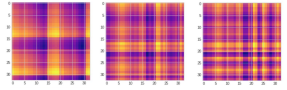
Transformer encoder SMILES



2D CNN Distance Matrix







Methods

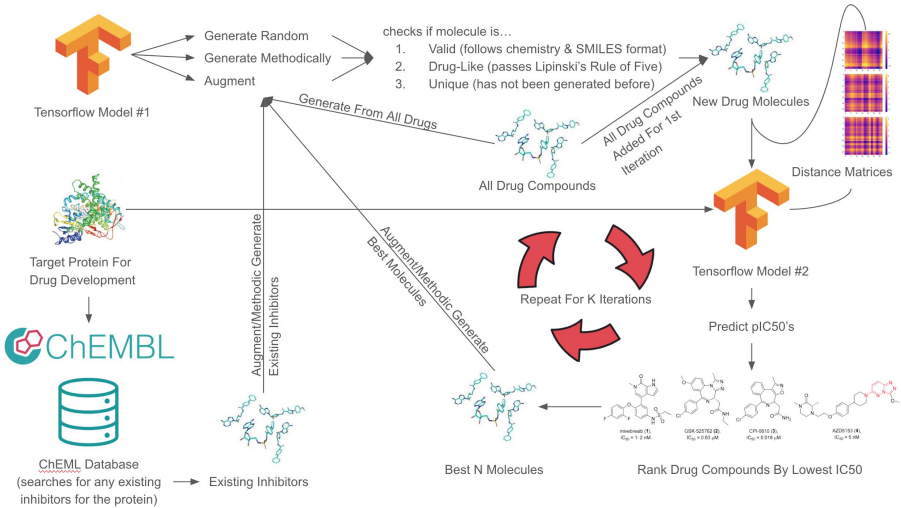
- Model #2 (CNN/Transformer neural network)
  - Trained on >500k protein-ligand interactions and validated on ~5k interactions
  - KIBA (kinase inhibitor bioactivity) Dataset
  - Weakly labelled data
    - Data included “< (value)” and “> (value)”
      - Removed “<,” “>,” “<=,” “>=”
  - Training Mean Absolute Error: ~0.76
  - Validation Mean Absolute Error: ~0.89

Methods

Putting It Together

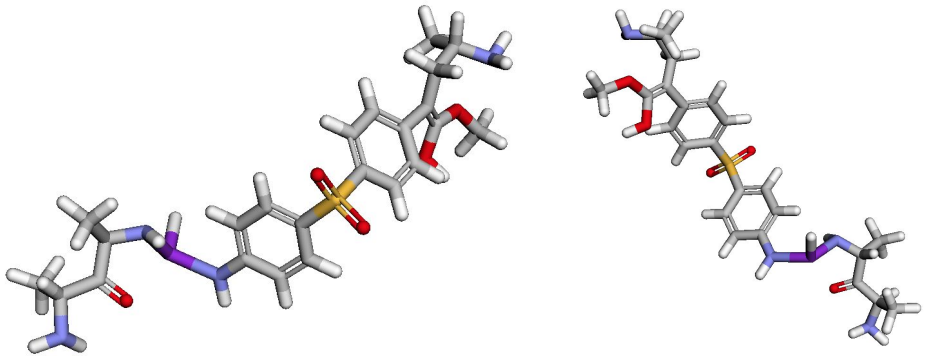
- Model #1 makes new drug-like compounds
- Model #2 predicts how good the compounds are
  - How well they bind to the target protein

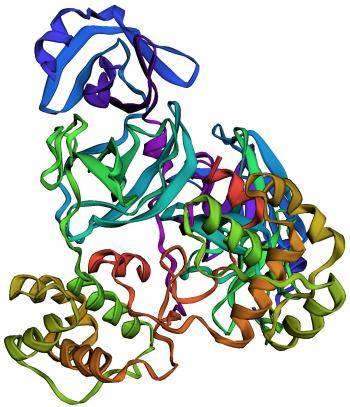
- Iterative process:
  - If existing inhibitors are present
    - Model #1 Augment/Generate Methodically on existing inhibitors
    - Model #2 Predict pIC50 value
    - Take best 50
    - Augment/Generate Methodically on best 50
    - Score
    - …
  - If no existing inhibitors are present
    - Model #1 Generate Randomly
    - Model #2 Predict pIC50 value
    - Take best 50
    - Augment/Generate Methodically on best 50
    - Score
    - …



Results

- Tried iterative process on the coronavirus (COVID) main protease (SARS 3C-like protease)
  - Has an essential role in processing the polyproteins that are translated from the viral RNA
  - Protein Sequence (length: 4382):
    - MESLVLGVNEKTHVQLSLPVLQVRDVLVRGFGDSVEEALSEAREHLKNGTCGLVELEKGVLPQLEQPYVFIKRSDALSTNHGHKVVELVAEMDGIQYGRSGITLGVLVPHVGETPIAYRNVLLRKNGNKGAGGHSYGIDLKSYDLGDELGTDPIEDYEQNWNTKHGSGALRELTRELNGGAVTRYVDNNFCGPDGYPLDCIKDFLARAGKSMCTLSEQLDYIESKRGVYCCRDHEHEIAWFTERSDKSYEHQTPFEIKSAKKFDTFKGECPKFV
FPLNSKVKVIQPRVEKKKTEGFMGRIRSVYPVASPQECNNMHLSTLMKCNHCDEVSWQTCDFLKATCEHCGTENLVIEGPTTCGYLPTNAVVKMPCPACQDPEIGPEHSVADYHNHSNIETRLRKGGRTRCFGGCVFAVYGCYNKRAYWVPRASADIGSGHTGITGDNVETLNEDLLEILSLRERVNINIVGDFHLNEEVAIILASFSASTSAFIDTIKSLDYKSFKTIVESCGNYKVTKGKPVKGAWNIGQQRSVLTPLCGFPSQAAGVIRSIFARTL
DAANHSPDLQRAAVTILDGISEQSLRLVDAMVYTSDLLTNSVIIMAYVTGGLVQQTSQWLSNLLGTTVEKLRPIFEWIEAKLSAGVEFLKDAWEILKFLIGVFDIVKGQIQVASDNIKDCVKCFIDVVNKALEMCIDQVTIAGAKLRSLNLGEVFIAQSKGLYRQCIRGKEQLQLLMPLKAPKEVTFLEGDSHDTVLTSEEVVLKNGELEALETPVDSFTNGAIVCTPYCVNGLMLLEWDKEGYCALSPGLLATNNVFRLKGGAPIKGVTFGEDTVWEV
QGYKNVRITFELDERVDKVLNEKCSVYTVESGTEVTEFACVVAEAVVKTLQPVSDLLTNMGIDLDEWSVATFYLFDDAGEENFSSRMYCSFYPPDEEEEDDAECEEEEIDETCEHEYGTEDDYQGLPLEFGASAETVRVEEEEEEDWLDDTTEQSEIEPEPEPTPEEPVNQFTGYLKLTDNVAIKCVDIVKEAQSANPMVIVNAANIHLKHGGGVAGALNKATNGAMQKESDDYIKLNGPLTVGGSCLLSGHNLAKKCLHVVGPNLNAGEDIQLL
KAAYENFNSQDILLAPLLSAGIFGAKPLQSLQVCVQTVRTQVYIAVNDKALYEQVVMDYLDNLKPRVEAPKQEEPPNTEDSKTEEKSVVVQKPVDVKPKIKACIDEVTTTLEETKFLTNKLLLFADINGKLYHDSQNMLRGEDMSFLEKDAPYMVGDVITSGDITCVVIPSKKAGGTTEMLSRALKKVPVDEYITTYPGQGCAGYTLEEAKTALKKCKSAFYVLPSEAPNAKEEILGTVSWNLREMLAHAEETRKLMPICMDVRAIMATIQRKYKGIKIQE
GIVDYGVRFFYTSKEPVASIITKLNSLNEPLVTMPIGYVTHGFNLEEAARCMRSLKAPAVVSVSSPDAVTTYNGYLTSSSKTSEEHFVETVSLAGSYRDWSYSGQRTELGVEFLKRGDKIVYHTLESPVEFHLDGEVLSLDKLKSLLSLREVKTIKVFTTVDNTNLHTQLVDMSMTYGQQFGPTYLDGADVTKIKPHVNHEGKTFFVLPSDDTLRSEAFEYYHTLDESFLGRYMSALNHTKKWKFPQVGGLTSIKWADNNCYLSSVLLALQQLEVK
FNAPALQEAYYRARAGDAANFCALILAYSNKTVGELGDVRETMTHLLQHANLESAKRVLNVVCKHCGQKTTTLTGVEAVMYMGTLSYDNLKTGVSIPCVCGRDATQYLVQQESSFVMMSAPPAEYKLQQGTFLCANEYTGNYQCGHYTHITAKETLYRIDGAHLTKMSEYKGPVTDVFYKETSYTTTIKPVSYKLDGVTYTEIEPKLDGYYKKDNAYYTEQPIDLVPTQPLPNASFDNFKLTCSNTKFADDLNQMTGFTKPASRELSVTFFPDLN
GDVVAIDYRHYSASPKKGAKLLHKPIVWHINQATTKTTFKPNTWCLRCLVWSTKPVDTSNSFEVLAVEDTQGMDNLACESQQPTSEEVVENPTIQKEVIECDVKTTEVVGNVILKPSDEGIKVTQELGHEDLMAAYVENTSITIKKPNELSLALGLKTIATHGIAAINSVPWSKLAYVKPFLGQAAITSNCAKRLAQRVFNNYMPYVFTLLFQLCTFTKSTNSRIRASLPTTIAKNSVKSVAKLCLDAGINYVKSPKFSKLFTIAMWLLLSICLCSSLICVT
AAFGVLLSNFGAPSYCNGVRELYLNSSNVTTMDFCEGSFPCSICLSGLDSLDSYPALETIQVTISSYKLDLTILGLAAEWVLAYMLFTKFFYLLGLSAIMQVFFGYFASHFISNSWLMWFIISIVQMAPVSAMVRMYIFFASFYYIWKSYVHIMDGCTSSTCMMCYKRNRATRVECTTIVNGMKRSFYVYANGGRGFCKTHNWNCLNCDTFCTGSTFISDEVARDLSLQFKRPINPTDQSSYIVDSVAVKNGALHLYFDKAGQKTYERHPLSHFVNLD
NLRANNTKGSLPINVIVFDGKSKCDESASKSASVYYSQLMCQPILLLDQALVSDVGDSTEVSVKMFDAYVDTFSATFSVPMEKLKALVATAHSELAKGVALDGVLSTFVSAARQGVVDTDVDTKDVIECLKLSHHSDLEVTGDSCNNFMLTYNKVENMTPRDLGACIDCNARHINAQVAKSHNVSLIWNVKDYMSLSEQLRKQIRSAAKKNNIPFRLTCATTRQVVNVITTKISLKGGKIVSTCFKLMLKATLLCVLAALVCYIVMPVHTLSIHDGYTN
EIIGYKAIQDGVTRDIISTDDCFANKHAGFDAWFSQRGGSYKNDKSCPVVAAIITREIGFIVPGLPGTVLRAINGDFLHFLPRVFSAVGNICYTPSKLIEYSDFATSACVLAAECTIFKDAMGKPVPYCYDTNLLEGSISYSELRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDAEYCRHGTCERSEVGICLSTSGRWVLNNEHYRALSGVFCGVDAMNLIANIFTPLVQPVGALDVSASVVAGGIIAILVTCAAYYFMKFRRVFGEYNHVVAANALLFL
MSFTILCLVPASYSFLPGVYSVFYLYLTFYFTNDVSFLAHLQWFAMFSPIVPFWITAIYVFCISLKHCHWFFNNYLRKRVMFNGVTFSTFEEEAALCTFLLNKEMYLKLRSETLLPLTQYNRYLALYNKYKYFSGALDTTSYREAACCHLAKALNDFSNSGADVLYQPPQTSITSAVLQSGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDTVYCPRHVICTAEDMLNPNYEDLLIRKSNHSFLVQAGNVQLRVIGHSMQNCLLRLKVDTSNPKTPK
YKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNHTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGKFYGPFVDRQTAQAAGTDTTITLNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCAALKELLQNGMNGRTILGSTILEDEFTPFDVVRQCSGVTFQGKFKKIVKGTHHWMLLTFLTSLLVQSTQWSLFFVYENAFLPFTLGIMAIAACAMLLVKHKHAFLCLFL
LPSLATVAYFNMVYMPASWVMRIMTWLELADTSLSGYRLKDCVMYASALVLILMTARTVYDDAARRVWTLMNVITLVYKVYYGNALDQAISMWALVISVTSNYSGVVTTIMFLARAIVFVCVEYYPLLFITGNTLQCIMLVYCFLGYCCCCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGLLPPKSSIDAFKLNIKLLGIGGKPCIKVATVQSKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVLLS
MQGAVDINRLCEEMLDNRATLQAIASEFSSLPSYAAYATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEFDRDAAMQRKLEKMADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALNNIINNARDGCVPLNIIPLTTAAKLMVVVPDYGTYKNTCDGNTFTYASALWEIQQVVDADSKIVQLSEINMDNSPNLAWPLIVTALRANSAVKLQNNELSPVALRQMSCAAGTTQTACTDDNALAYYNNSKGGRFVLALLSDHQDLKW
ARFPKSDGTGTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDPAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLRNTVCTVCGMWKGYGCSCDQLREPLMQSADASTFLNGFAV
  - Best Generated Drug Candidate: “CC(C(=O)C(C)N)NINc1ccc(S(=O)(=O)c2ccc(C(CC(C)N)=C(O)OC)cc2)cc1”

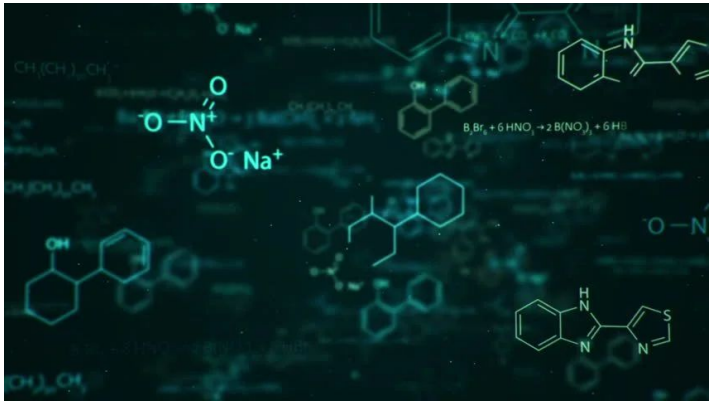




Results

The process definitely met my engineering goals

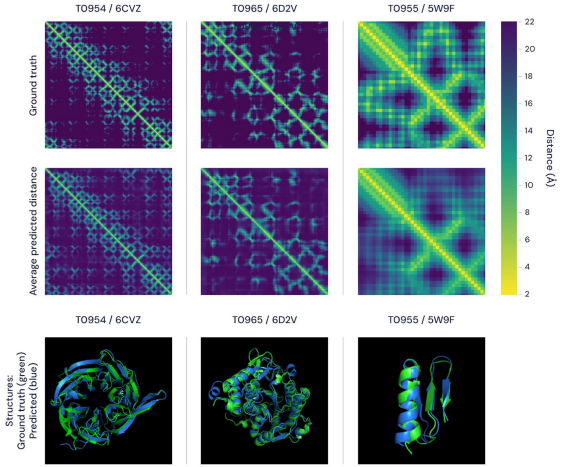
- Model #1 is able to generate valid, drug-like compounds ~82% of the time, with an overall training accuracy of ~90.1% and validation accuracy of ~89.2%
- Model #2 is able to accurately and quickly predict the pIC50 of a drug-target interaction, achieving a training mean absolute error of ~0.76 and validation mean absolute error of ~0.89
- Both models are able to generate and score tens of thousands of compounds in a day and are able to create drug-candidates with a higher predicted binding affinity than existing inhibitors for the SARS 3C-like protease



Discussion

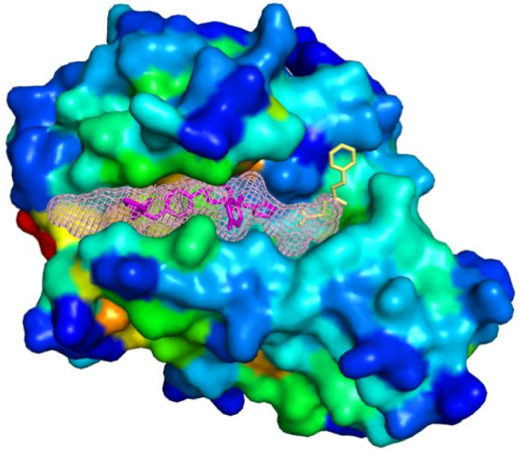
- What do these results mean?
  - The results prove that the models have the potential to be production-ready for use in the drug discovery process

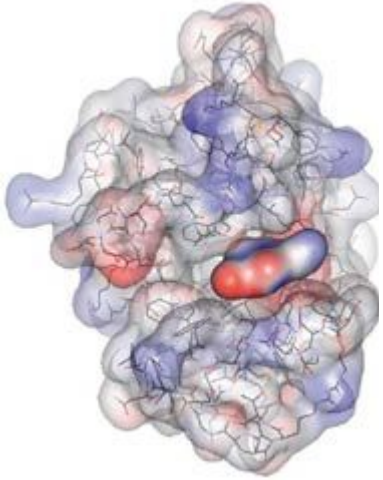
- Problems?
  - There were definitely a lot of unexpected issues, but I simply edited the code to avoid them the next time I run the program
  - The major problem that could not necessarily be fixed with edited code was having very weakly-labelled data for the second model
  - The model would have to learn very complex 3D interactions from this weakly-labelled data
    - Protein folding (predicting the 3D structure of a protein) is one of the hardest problems and has only very recently been solved

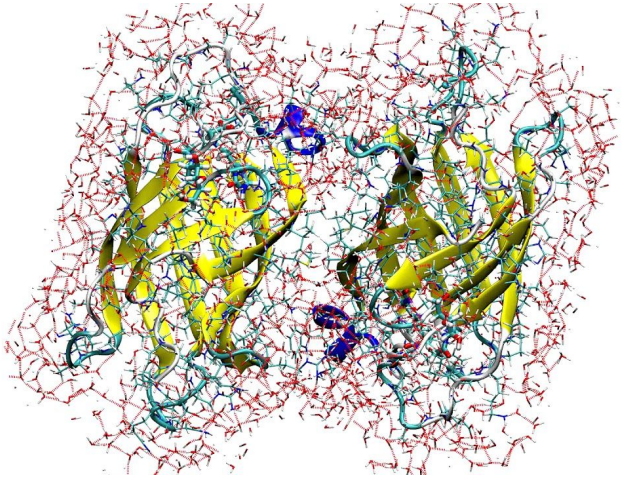


Discussion

- How is my prototype an improvement or advancement over what is currently available? (physics-based computer simulations)
  - The prototype is far less computationally expensive than physics-based computer simulations
  - The prototype can also generate new drug compounds instead of only testing existing ones
  - The prototype would work for new viruses which have unsolved target protein structures



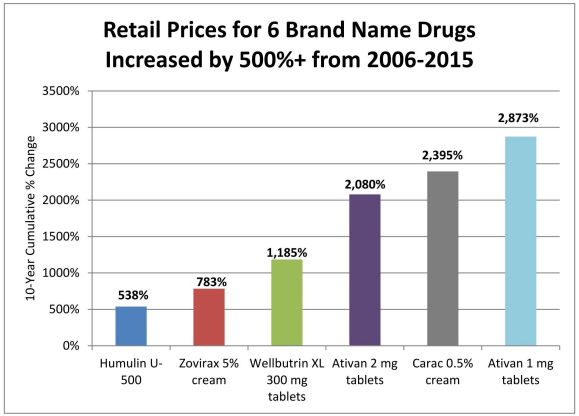


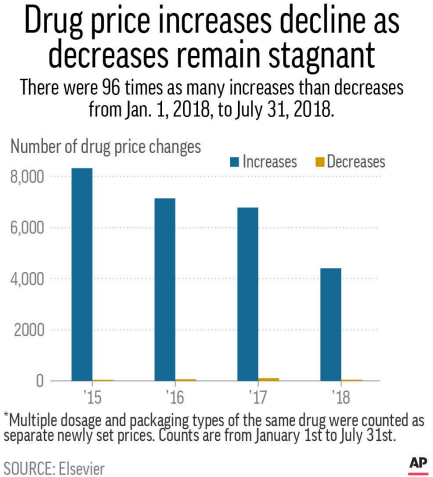


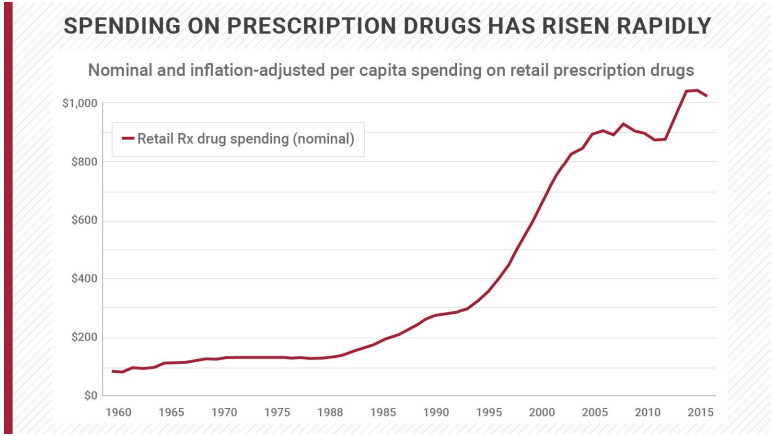
Conclusions

Applications:

- If these models were upgraded and trained with more data, the process could be production-ready for pharmaceutical scientists to utilize in the drug discovery process
- Drugs could be made much faster and with less expense
  - This would allow drugs to be cheaper as well, which is great since prices are drastically increasing







References

Benet, Leslie Z, et al. “BDDCS, the Rule of 5 and Drugability.” *Advanced Drug Delivery Reviews*, U.S. National Library of Medicine, 1 June 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4910824/.

Commissioner, Office of the. “The Drug Development Process.” *U.S. Food and Drug Administration*, FDA, www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.

Gaulton, Anna, et al. “ChEMBL: a Large-Scale Bioactivity Database for Drug Discovery.” *Nucleic Acids Research*, Oxford University Press, Jan. 2012, www.ncbi.nlm.nih.gov/pmc/articles/PMC3245175/.

Mohs, Richard C, and Nigel H Greig. “Drug Discovery and Development: Role of Basic Biological Research.” *Alzheimer's & Dementia (New York, N. Y.)*, Elsevier, 11 Nov. 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5725284/.

Paul, Debleena, et al. “Artificial Intelligence in Drug Discovery and Development.” *Drug Discovery Today*, Elsevier Ltd., Jan. 2021, www.ncbi.nlm.nih.gov/pmc/articles/PMC7577280/#:~:text=AI%20can%20be%20used%20effectively,and%20uncertainty%20of%20the%20data.