Target2Drug : a novel programmatic workflow to automate *In Silico* drug discovery

Ben Geoffrey A S*[a,], Rafal Madaj[b], Akhil Sanker[c], Pavan Preetham Valluri[d] , Judith Gracia[a] , Harshmeet Singh[e]

[a] *University of Madras, Chepauk, Chennai 600 005, India*
[b] *Centre of Molecular and Macromolecular Studies, Polish Academy of Sciences, Poland*
[c] *SRM University, Tamil Nadu 603203, India*
[d] *PSG College of Technology, Coimbatore, India*
[e] *Independent Researcher, India*

*Corresponding author email : bengeof@gmail.com

Abstract

As the Big Data and Artificial Intelligence (AI) revolution continues to affect every area of our lives, it's influence is also exerted in the areas of bioinformatics, computational biology and drug discovery. Machine/Deep Learning tools have been developed to predict compounds-drug target interactions and the vice-versa process of predicting target interactions for an compound. In our presented work, we report a programmatic tool, which incorporates many features of the bioinformatics, computational biology and AI-driven drug discovery revolutions into a single workflow assembly. When a user is required to identify drugs against a new drug target, the user provides target signatures in the form of amino acid sequence of the target or it's corresponding nucleotide sequence as input to the tool and the tool carries out a BLAST protocol to identify known protein drug targets that are similar to the new target submitted by the user and collects data linked to the target involving, active compounds against the target, the activity value and molecular descriptors of active compounds to perform QSAR modelling and to generate drug leads with predictions from the validated QSAR model. The tool performs an *In-Silico* modelling to generate *In-Silico* interaction profiles of compounds generated as drug leads and the target and stores the results in the working folder of the user. To demonstrate the use of the tool, we have carried out a demonstration with the target signatures of the current pandemic causing virus, SARS-CoV 2. However the tool can be used against any target and is expected to help in growing our knowledge graph of targets and interacting compounds.

The program is hosted, maintained and supported at the GitHub repository link given below

https://github.com/bengeof/Target2DrugChemRxivNotebook

Introduction

The Big Data and Artificial Intelligence revolution is leaving its mark on every sphere of human life and research is no exception. The sphere of bioinformatics, computational biology and drug discovery has been approached with many data-driven and AI based approaches[1-9]. The UNIPROT, RCSB database, GenBank database and the PubChem database provide the required big data in the areas of proteomics, genomics and small drug molecule discovery to employ machine/deep learning methods to these areas [10-13]. In small drug molecule discovery, the recent surge in development of deep learning based protein-ligand interaction tools helps has helped researchers in identifying small drug molecules that can interact with a particular drug target [14-16]. One observes two complimentary approaches of deep learning tools in small drug molecule discovery. Deep learning tools have been developed to predict drug targets any known compound can interact with while also the other complimentary approach of predicting the compounds that can interact with a particular drug target has also been reported in literature by research groups. The advantage of using deep learning tools over the previously used *In Silico* modelling in identifying small drug molecules and predicting compound-target interactions is that, deep learning based compound-target interaction identification is carried out at a much lesser computational cost as compared to *In Silico* modelling. Wang, Y. B. et al have developed a deep learning based tool based on the LSTM neural network architecture to predict drug-target interaction [17]. Wen, M. Et al also report a deep learning based tool for drug-target prediction [18]. Geoffrey AS et al have developed a hybrid approach for predicting compound-drug target and drug target-compound interactions which attempt to combine advantages of both deep learning and *In Silico* modelling based approaches in a singular tool without overshooting the computational expense [19, 20]. Similarly machine/deep learning methods have been used on proteomic and genomic data to identify bio-markers and drug targets [21]. DG IJzendoorn et al have used machine/deep learning methods to identify cancer biomarkers and novel therapeutic targets [22]. In our presented work, we propose a programmatic workflow, which incorporates many features of these independent developments related to machine/deep learning driven drug discovery into a single workflow assembly which makes for a more wholistic and automated data-driven drug discovery workflow. When required to identify drugs against a new drug target, the user of the tool may provide target signatures in the form of amino acid sequence of the target or it’s corresponding nucleotide sequence as input to the tool and the programmatic workflow to identify drugs that can be used against the target proceeds as follows. The tool carries out a BLAST protocol with the new target signatures provided by the user and identifies known protein drug targets that are similar to the new target submitted by the user. Among the know protein drug targets that are identified by the BLAST protocol of the tool, the tool identifies targets with data availability on PubChem to perform QSAR based drug lead generation. In order to perform QSAR based drug lead generation, the tool collects the data of reported experimental inhibition activity of PubChem compounds against the target and the molecular descriptors of the active compounds to perform QSAR modelling. A machine learning based AutoQSAR protocol of training, validation and prediction was carried out for drug lead generation, to generate drug leads from the huge ligand library of PubChem. To perform *In Silico* modelling of the interaction of the compounds generated by the tool as drug leads and the protein drug target, a popular high throughput virtual screening package AutoDock-Vina was used programmatically through the tool. The protein-ligand interaction profiles are generated and results are stored in the working folder of the user. A detailed methodology and the demonstrated use of the tool with COVID-19 target signatures can be found below. While a demonstration is carried out with the target signatures of the virus causing the present pandemic, the tool can be used against any target and is expected to help in increasing our knowledge graph of targets and interacting compounds.

Materials and methods

The algorithmic workflow of the tool pictorial represented by way of a block diagram is shown in Fig. 1. To detail the algorithmic workflow of the tool by way of a demonstrated example

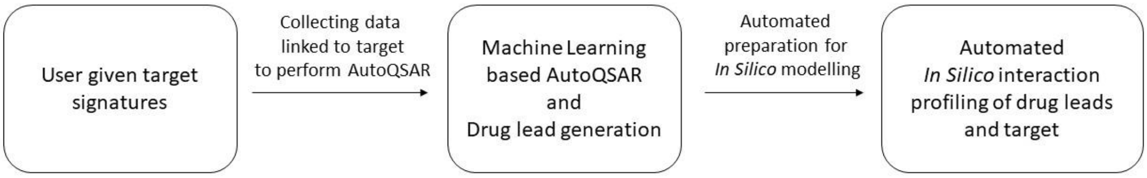
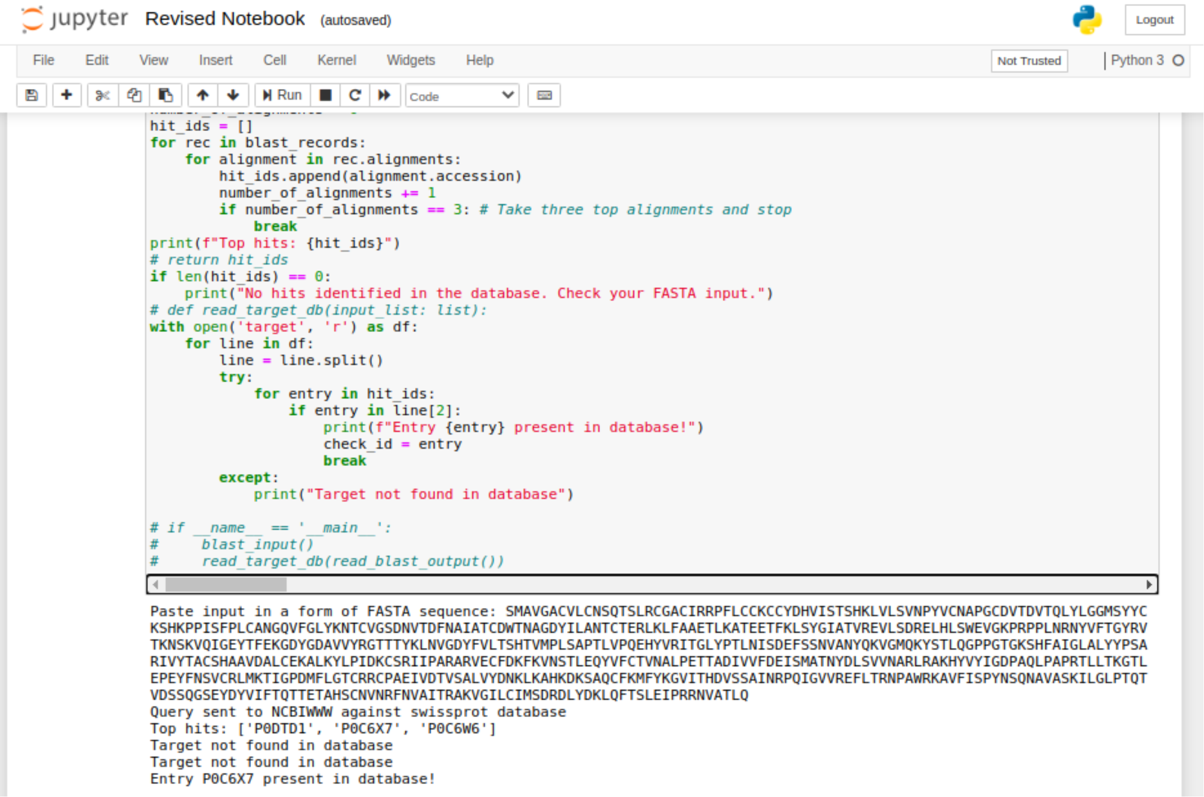

User given target signatures
Collecting data linked to target to perform AutoQSAR
Machine Learning based AutoQSAR and Drug lead generation
Automated preparation for *In Silico* modelling
Automated *In Silico* interaction profiling of drug leads and target

Fig.1 – Algorithmic workflow

we have chosen the COVID-19 target signatures as input given to the tool in view of the pandemic situation caused by the virus presently. Target signatures of amino acid sequence of the SARS-CoV 2 proteome with NCBI reference of NC_045512 was given as input to the tool as shown in Fig.2



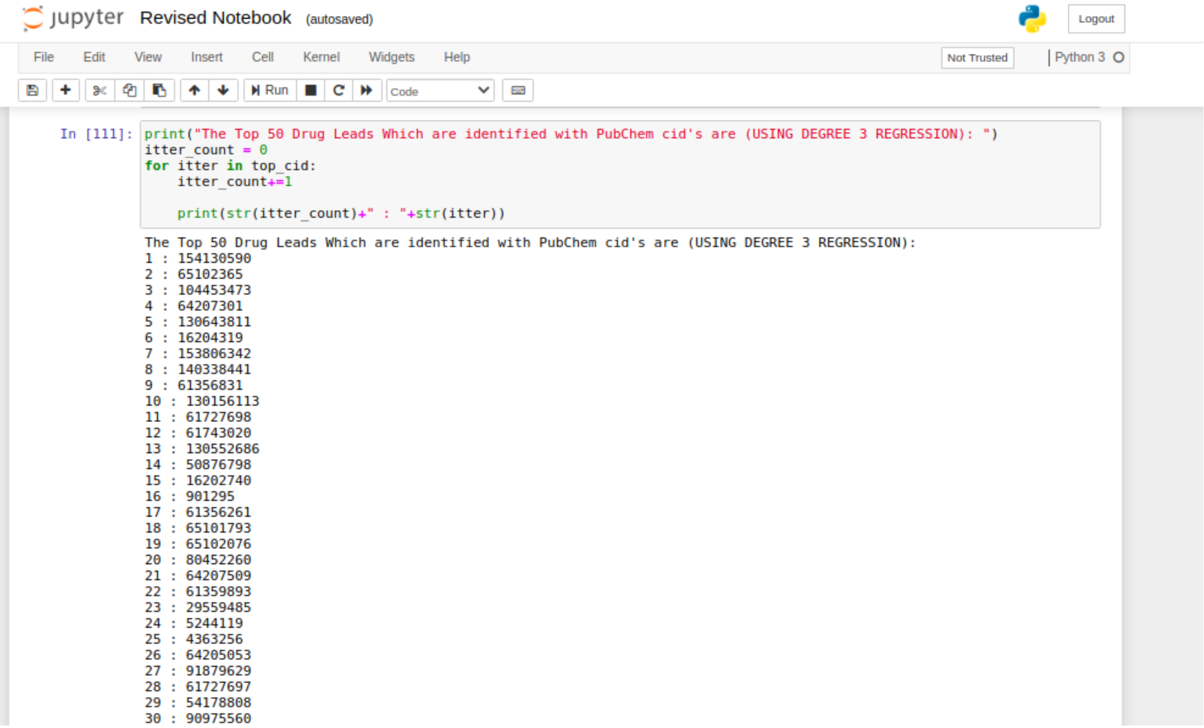
```
hit_ids = []
for rec in blast_records:
    for alignment in rec.alignments:
        hit_ids.append(alignment.accession)
        number_of_alignments += 1
        if number_of_alignments == 3: # Take three top alignments and stop
            break
print(f"Top hits: {hit_ids}")
# return hit_ids
if len(hit_ids) == 0:
    print("No hits identified in the database. Check your FASTA input.")
# def read_target_db(input_list: list):
with open('target', 'r') as df:
    for line in df:
        line = line.split()
        try:
            for entry in hit_ids:
                if entry in line[2]:
                    print(f"Entry {entry} present in database!")
                    check_id = entry
                    break
        except:
            print("Target not found in database")

# if __name__ == '__main__':
#     blast_input()
#     read_target_db(read_blast_output())
```

```
Paste input in a form of FASTA sequence: SMAVGACVLCNSQTSLRCGACIRRPFLCCKCCYDHVISTSHKLVLSVNPYVCNAPGCDVTDVTQLYLGGMSYYC
KSHKPPISFPLCANGQVFGLYKNTCVGSDNVTDFNAIATCDWTNAGDYILANTCTERLKLFAAETLKATEETFKLSYGIATVREVLSDRELHLSWEVGKPRPPLNRNYVFTGYRV
TKNSKVQIGEYTFEKGDYGDAVVYRGTTTYKLNVGDYFVLTSHTVMPLSAPTLVPQEHYVRITGLYPTLNISDEFSSNVANYQKVGMQKYSTLQGPPGTGKSHFAIGLALYYPSA
RIVYTACSHAAVDALCEKALKYLPIDKCSRIIPARARVECFDKFKVNSTLEQYVFCTVNALPETTADIVVFDEISMATNYDLSVVNARLRAKHYVYIGDPAQLPAPRTLLTKGTL
EPEYFNSVCRLMKTIGPDMFLGTCRRCPAEIVDTVSALVYDNKLKAHKDKSAQCFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVASKILGLPTQT
VDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCIMSDRDLYDKLQFTSLEIPRRNVATLQ
Query sent to NCBIWWW against swissprot database
Top hits: ['P0DTD1', 'P0C6X7', 'P0C6W6']
Target not found in database
Target not found in database
Entry P0C6X7 present in database!
```

Fig.2 – collection of target signatures from user

After collecting the target signatures in the form of amino acid sequence of target or it's corresponding nucleotide sequence, the tool performs a BLAST [23] protocol to identify UNIRPOT target ID's which are similar to input target signatures. Among the identified UNIPROT ID's which are similar the input target signature, the tool identifies UNIPROT target ID's with availability of data on PubChem required to perform AutoQSAR [24]. The identified data of active compounds against the target, the experimental activity value and molecular descriptors of active compounds is fetched from PubChem by the program through WebAPI programmatic access to PubChem [25]. The tool performs a machine learning based AutoQSAR protocol which involves training, validation and prediction. QSAR models are usually linear or non-linear statistical correlation between the experimental activity and molecular descriptors. While the total number of descriptors is 8, the program builds a QSAR model with every possible combination of descriptors by generating all possible combinations of descriptors where n = 8 and r = 2, 3, 4, 5, 6, 7 and $^{n}C_{r}$ in such a case gives a total of 256 combinations of descriptor selection for the QSAR model. The program builds a linear and non-linear regression based QSAR model with all 256 possible combinations of descriptors and selects the QSAR with highest R^2 value or R^2 value closest to 1. The model with R^2 value closest to 1 is chosen for prediction. The prediction is carried with the large chemical library of PubChem compounds that are structurally associative to the compounds active against the target. The program prints out the top 50 compounds of the prediction as drug leads against the target as shown in Fig.3 and the drug leads are required to satisfy the Lipinski's drug likeness criteria. The tool, programmatically fetches the structures of compounds predicted as drug leads against the target and the structure of the target and prepares them for molecular docking. The ligand and receptor preparation is carried out with standard AutoDockTools(ADT) scripts. AutoDock-Vina is run programmatically through the tool, and an *In-Silico* interaction profile of compound and drug target are stored in the working folder of the user [26].



In [111]:

```
print("The Top 50 Drug Leads Which are identified with PubChem cid's are (USING DEGREE 3 REGRESSION): ")
itter_count = 0
for itter in top_cid:
    itter_count+=1

    print(str(itter_count)+" : "+str(itter))
```

```
The Top 50 Drug Leads Which are identified with PubChem cid's are (USING DEGREE 3 REGRESSION):
1 : 154130590
2 : 65102365
3 : 104453473
4 : 64207301
5 : 130643811
6 : 16204319
7 : 153806342
8 : 140338441
9 : 61356831
10 : 130156113
11 : 61727698
12 : 61743020
13 : 130552686
14 : 50876798
15 : 16202740
16 : 901295
17 : 61356261
18 : 65101793
19 : 65102076
20 : 80452260
21 : 64207509
22 : 61359893
23 : 29559485
24 : 5244119
25 : 4363256
26 : 64205053
27 : 91879629
28 : 61727697
29 : 54178808
30 : 90975560
```

Fig. 3 – Drug leads identified by the tool

Running the program requires no more programming knowledge than running the python executable file in python 3 environment in Linux OS along with some python dependency packages installed such as:

pandas

biopandas

numpy

matplotlib

scikit-learn

seaborn

selenium (along with selenium's driver for firefox browser)

Other additional dependencies for automated *In Silico* modelling

openbabel 2.4.1

mgltools 1.5.4

autodock-vina 1.1.2-4

The program is hosted, maintained and supported at the GitHub repository link given below
https://github.com/bengeof/Target2DrugChemRxivNotebook

Results and discussion

For the user given target signatures, the program identifies drug leads that satisfy Lipinski's drug likeness criteria from the large ligand library of PubChem database. For the top 50 drug leads the program automatically perform *In-Silico* modelling of compound-target interaction and stores the *In-Silico* generated interaction profiles in the working folder of the user. The top 50 drug leads and their *In-Silico* interaction scores from AutoDock-Vina are given in Table 1.

PubChem CID	Target	Binding energy
142747435	PDB ID : 1qz8	-6.8
16203797		-6.5
1580642		-6.2
16075059		-6.2
91879629		-6.2
142747432		-6.2
62024579		-6.1
61727697		-6
61743020		-6

62024757		-6
2196453		-5.9
61360057		-5.9
61727698		-5.9
91875621		-5.9
44589253		-5.7
61356831		-5.7
61359893		-5.6
4363256		-5.5
70485909		-5.5
901295		-5.4
3542734		-5.4
16202740		-5.4
61356261		-5.4
130156113		-5.4
75268360		-5.3
140338441		-5.3
150985451		-5.3
154130590		-5.3
29559485		-5.2
68862352		-5.2
153806342		-5.1
16203618		-5
16204319		-5
47391569		-5
50876798		-5
80452260		-4.9
5244119		-4.8
65102076		-4.8
90975560		-4.7
104453473		-4.7
54178808		-4.6
130552686		-4.6
153793082		-4.6
64207301		-4.5
65101793		-4.5
65102365		-4.5
130643811		-4.5
64205053		-4.4
65237247		-4.3
64207509		-4.2

Select visualization of drug candidate and target interaction is shown in Fig.3.

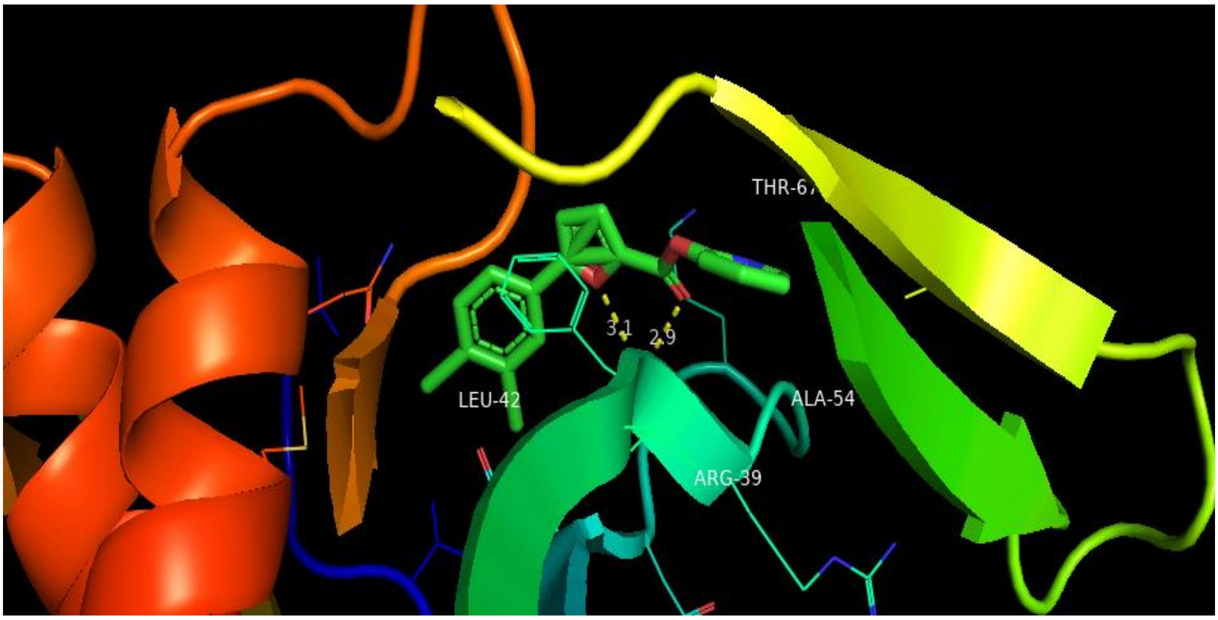


Fig. 3 – Interaction of compound with PubChem ID 142747435 and target with PDB ID 1qz8

The results indicate that the generated drug leads interact favourably with the target and thus making a case of the use of tool which helps in automated small drug molecule candidate identification when the target signatures are provided by the user.

Conclusion

In our presented work, we report a programmatic tool, which incorporates many features of the bioinformatics, computational biology and AI-driven drug discovery revolutions into a single workflow assembly. When a user is required to identify drugs against a new drug target, the user provides target signatures in the form of amino acid sequence of the target or it's corresponding nucleotide sequence as input to the tool and the tool carries out a BLAST protocol to identify known protein drug targets that are similar to the new target submitted by the user and collects data linked to the target involving active compounds against the target, the activity value and molecular descriptors of active compounds to perform QSAR modelling and drug lead generation with predictions based on the validated QSAR model. The tool performs an *In-Silico* modelling to generate *In-Silico* interaction profiles of compounds generated as drug leads and the target and stores the results in the working folder of the user. To demonstrate the use of the tool, we have carried out a demonstration with the target signatures of the current pandemic causing virus, SARS-CoV 2. The results indicated that the generated drug leads interacted favourably with the target and thus making a case for the use of tool which helps in automated small drug molecule candidate identification when the drug target signatures are provided by the user. The tool can be used in the case of any target and is expected to help in growing our knowledge graph of targets and interacting compounds.

References

1. Min, Seonwoo, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics." *Briefings in bioinformatics* 18, no. 5 (2017): 851-869.
2. Li, Yu, Chao Huang, Lizhong Ding, Zhongxiao Li, Yijie Pan, and Xin Gao. "Deep learning in bioinformatics: Introduction, application, and perspective in the big data era." *Methods* 166 (2019): 4-21.
3. Lan, Kun, Dan-tong Wang, Simon Fong, Lian-sheng Liu, Kelvin KL Wong, and Nilanjan Dey. "A survey of data mining and deep learning in bioinformatics." *Journal of medical systems* 42, no. 8 (2018): 139.
4. Cao, Yue, Thomas Andrew Geddes, Jean Yee Hwa Yang, and Pengyi Yang. "Ensemble deep learning in bioinformatics." *Nature Machine Intelligence* (2020): 1-9.
5. Li, Haoyang, Shuye Tian, Yu Li, Qiming Fang, Renbo Tan, Yijie Pan, Chao Huang, Ying Xu, and Xin Gao. "Modern deep learning in bioinformatics." *Journal of molecular cell biology* (2020).
6. Jones, William, Kaur Alasoo, Dmytro Fishman, and Leopold Parts. "Computational biology: deep learning." *Emerging Topics in Life Sciences* 1, no. 3 (2017): 257-274.
7. Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. "Deep learning for computational biology." *Molecular systems biology* 12, no. 7 (2016): 878.
8. Goh, Garrett B., Nathan O. Hodas, and Abhinav Vishnu. "Deep learning for computational chemistry." *Journal of computational chemistry* 38, no. 16 (2017): 1291-1307.
9. Sastry, Anand, Jonathan Monk, Hanna Tegel, Mathias Uhlen, Bernhard O. Palsson, Johan Rockberg, and Elizabeth Brunk. "Machine learning in computational biology to accelerate high-throughput protein expression." *Bioinformatics* 33, no. 16 (2017): 2487-2495.
10. UniProt Consortium. "UniProt: a hub for protein information." *Nucleic acids research* 43, no. D1 (2015): D204-D212.
11. Rose, Peter W., Bojan Beran, Chunxiao Bi, Wolfgang F. Bluhm, Dimitris Dimitropoulos, David S. Goodsell, Andreas Prlić et al. "The RCSB Protein Data

Bank: redesigned web site and web services." *Nucleic acids research* 39, no. suppl_1 (2010): D392-D401.

12. Benson, Dennis A., Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and Eric W. Sayers. "GenBank." *Nucleic acids research* 41, no. D1 (2012): D36-D42.
13. Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han et al. "PubChem substance and compound databases." *Nucleic acids research* 44, no. D1 (2016): D1202-D1213.
14. Stepniewska-Dziubinska, Marta M., Piotr Zielenkiewicz, and Pawel Siedlecki. "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction." *Bioinformatics* 34, no. 21 (2018): 3666-3674.
15. Colwell, Lucy J. "Statistical and machine learning approaches to predicting protein–ligand interactions." *Current opinion in structural biology* 49 (2018): 123-128.
16. Deng, Wei, Curt Breneman, and Mark J. Embrechts. "Predicting protein− ligand binding affinities using novel geometrical descriptors and machine-learning methods." *Journal of chemical information and computer sciences* 44, no. 2 (2004): 699-703.
17. Wang, Y. B., You, Z. H., Yang, S., Yi, H. C., Chen, Z. H., & Zheng, K. (2020). A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Medical Informatics and Decision Making*, *20*(2), 1-9
18. Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., & Lu, H. (2017). Deep-learning-based drug–target interaction prediction. *Journal of proteome research*, *16*(4), 1401-1409
19. Geoffrey AS, Ben, Pavan Preetham Valluri, Akhil Sanker, Rafal Madaj, Host Antony Davidd, Beutline Malgija, Konka Dinesh et al. "Compound2Drug–a Machine/deep Learning Tool for Predicting the Bioactivity of PubChem Compounds." (2020).
20. Geoffrey A S, Ben; Madaj, Rafal; Sanker, Akhil; Tresanco, Mario Sergio Valdés; Davidd, Host Antony; Roy, Gitanjali; et al. (2020): Automated In Silico Identification of Drug Candidates for Coronavirus Through a Novel Programmatic Tool and Extensive Computational (MD, DFT) Studies of Select Drug Candidates. ChemRxiv. Preprint.
21. Zhang, Z., & Liu, Z. P. (2019, August). Identifying Cancer Biomarkers from High-Throughput RNA Sequencing Data by Machine Learning. In *International Conference on Intelligent Computing* (pp. 517-528). Springer, Cham.

22. van IJzendoorn, D. G., Szuhai, K., Briaire-de Bruijn, I. H., Kostine, M., Kuijjer, M. L., & Bovée, J. V. (2019). Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS computational biology, 15*(2), e1006826.
23. Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic acids research* 25, no. 17 (1997): 3389-3402.
24. Dixon, Steven L., Jianxin Duan, Ethan Smith, Christopher D. Von Bargen, Woody Sherman, and Matthew P. Repasky. "AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling." *Future medicinal chemistry* 8, no. 15 (2016): 1825-1839.
25. Kim, S., Thiessen, P. A., Bolton, E. E., & Bryant, S. H. (2015). PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. Nucleic Acids Research, 43(W1). doi:10.1093/nar/gkv396
26. Huey, Ruth, Garrett M. Morris, and Stefano Forli. "Using AutoDock 4 and AutoDock Vina with AutoDockTools: A Tutorial." *The Scripps Research Institute Molecular Graphics Laboratory* (2012).