A programmatic tool for automatic ease in coronavirus drug discovery through programmatically automated data mining, QSAR and *In Silico* modelling

Ben Geoffrey A S*[a], Rafal Madaj[b], Akhil Sanker[c] Mario Sergio Valdés Tresanco[d], Host Antony David [e], Gitanjali Roy[f]

[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] University of Medellin, Medellin, Colo mbia
[e] Madras Christian College, Chennai 600 059, India
[f] Indiana University School of Medicine, Indianapolis, USA

*Corresponding author email : bengeof@gmail.com

Abstract

The work is composed of python based programmatic tool that automates the workflow of drug discovery for coronavirus. Firstly, the python program is written to automate the process of data mining PubChem database to collect data required to perform a machine learning based AutoQSAR algorithm through which drug leads for coronavirus are generated. The data acquisition from PubChem was carried out through python web scrapping techniques. The workflow of the machine learning based AutoQSAR involves feature learning and descriptor selection, QSAR modelling, validation and prediction. The drug leads generated by the program are required to satisfy the Lipinski's drug likeness criteria as compounds that satisfy Lipinski's criteria are likely to be an orally active drug in humans. Drug leads generated by the program are fed as programmatic inputs to an In Silico modelling package to computer model the interaction of the compounds generated as drug leads and two coronavirus drug targets identified with their PDB ID : 6W9C and 1P9U. The results are stored in the working folder of the user. The program also generates protein-ligand interaction profiling and stores the visualized images in the working folder of the user. Thus our programmatic tool ushers in the new age automatic ease in drug identification for coronavirus through a fully automated QSAR and an automated In Silico modelling of the drug leads generated by the autoQSAR algorithm.

Introduction

PubChem is a data repository of chemical compounds, their properties and biological activities [1] which can be programmatically accessed through web API packages such as PUB-REST and python web scrapping techniques [2,3]. Quantitative Structure-Activity Relationship(QSAR) studies are statistical based studies through which drug leads are generated which provide cost cutting advantages in testing and drug discovery for the pharmaceutical industry [4-7]. However the data set required to perform a QSAR study is curated by researchers before performing the statistically study. Our programmatic tool automates this process of data acquisition required to perform a QSAR study to generate drug leads for coronavirus through programmatic access of PubChem database and python web scrapping techniques [8,9]. The workflow of the QSAR study was also automated through a machine learning based AutoQSAR algorithm. The workflow of a machine learning based AutoQSAR algorithm involves feature learning and descriptor selection, QSAR modelling, validation and prediction [10-12]. The drug leads generated by the program are required to satisfy the Lipinski's drug likeness criteria as compounds that satisfy Lipinski's criteria are likely to be an orally active drug in humans [13]. Drug leads generated by the program are fed as programmatic inputs to an In Silico modelling package to computer model the interaction of the compounds generated as drug leads and two coronavirus drug targets identified with their PDB ID : 6W9C and 1P9U. The coronavirus drug targets identified from the literature were the main protease of coronavirus (PDB ID : 1P9U) and papain like protease (PDB ID :6W9C) [14,15]. The results of the *In Silico* modelling are stored in the working folder of the user. The program also generates protein-ligand interaction profiling and stores the visualized images in the working folder of the user. Thus our programmatic tool ushers in the new age automatic ease in drug identification for coronavirus through a fully automated QSAR and an automated In Silico modelling of the drug leads generated by the autoQSAR algorithm.   .

Our work is distinguished from previous attempts of virtual screening of large ligand libraries in way that we employ programmatic techniques as compared to other works that do not [16]. However as compared to recent data drive machine learning based approaches to drug discovery [17,18] that use pre-downloaded data sets we deploy a real time data mining which makes a case for a dynamic approach to drug lead generation and the results of the program are reflective of PubChem data library at the instant the program is run and thus

approaches drug discovery from a dynamic approach in an age Big Data and constantly growing data libraries. We also add to the existing richness of the novelty of methods [18-20] in data driven drug discovery in the following way. Our programmatic tool couples the drug leads generated by the AutoQSAR algorithm as programmatic inputs to an *In Silico* modelling package and programmatically profiles the protein-ligand interaction and stores the results in the working folder of the user. While adding to the existing richness in data driven machine learning based drug discovery methods, our work also adds new scientific findings to existing literature as the two coronavirus targets chosen for the study have been approached by data driven machine learning based methods.

Methods and Techniques

The workflow of the programmatic tool implementing programmatic data mining, AutoQSAR and automated In Silico modelling for identification of drugs against coronavirus is shown in Fig.1. The first process involved in programmatic workflow is the data mining of PubChem database to automate the process of data acquisition to implement a machine learning based AutoQSAR algorithm that automates the process of drug lead generation for coronavirus. The programmatic access to PubChem is accomplished through python commands [8, 9]. The next process in the workflow involves implementing the machine learning based AutoQSAR algorithm for drug lead generation. The drug leads are generated by the AutoQSAR algorithm through the workflow that involves feature learning, descriptor selection, QSAR modelling, validation and prediction [10-12]. Based on the validated QSAR model PubChem compound library is screened for drug lead generation based on the validated QSAR model. The drug leads generated by the program are filtered through the Lipinski's criteria [13].

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

Data mining of PubChem Database

1. Query PubChem to collect data to perform AutoQSAR

2. Generate drug leads using a machine learning based AutoQSAR algorithm

Drug lead generation
(Machine Learning AutoQSAR)

3. Apply Lipinski's rule of drug likeness to select drug like compounds

4. Drug leads as programmatic inputs to perform automated *In Silico* modelling

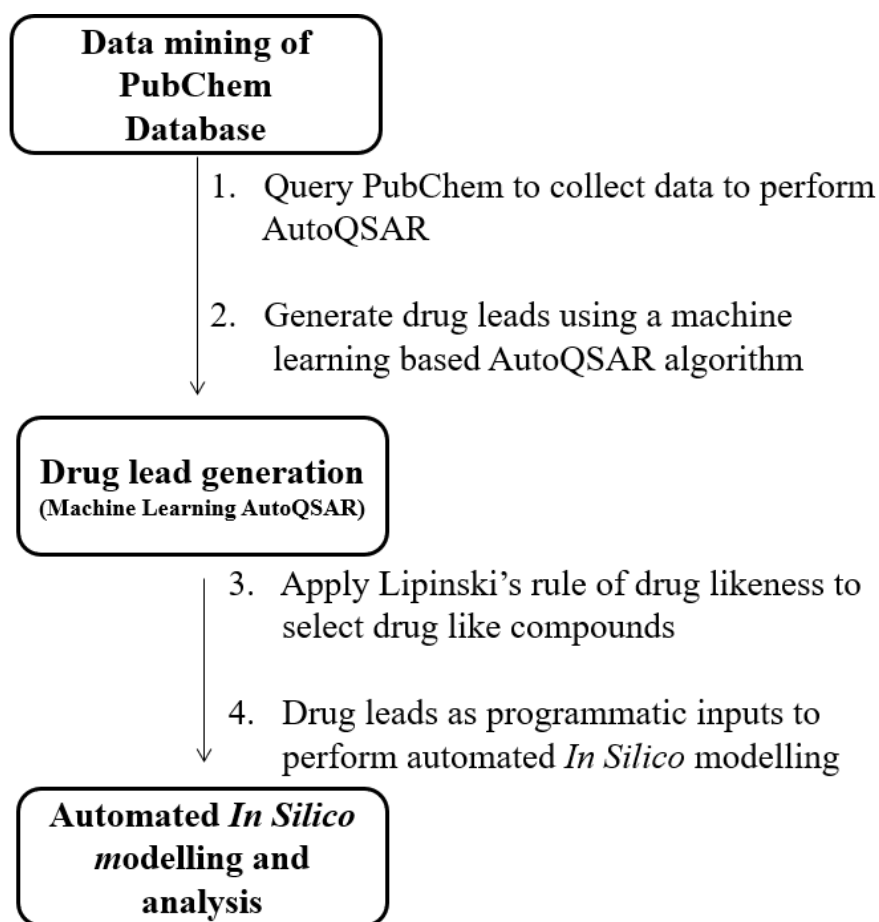**Automated *In Silico* modelling and analysis**

Fig. 1 – Workflow of the programmatic tool to automate the drug discovery for coronavirus

The running of the program requires a stable internet connection and the run time of the program is expected to be a few hours however it is expected to vary based on CPU and internet speed. The program prints out the PubChem CIDs of top 30 compounds identified as drug leads for coronavirus drugs (shown in Fig.2).

```
for itter in top_cid:
    itter_count+=1

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

The Top 30 Drug Leads Which are identified with PubChem cid's are:
1 : 17154971
2 : 1909002
3 : 1713563
4 : 17154975
5 : 8812727
6 : 17154974
7 : 214600
8 : 1581092
9 : 3542734
10 : 1581052
11 : 8811193
12 : 273043
13 : 8812498
14 : 1645090
15 : 12480412
16 : 12480411
17 : 316076
18 : 57525
19 : 2485698
20 : 44393555
21 : 16358
22 : 8811804
23 : 1713771
24 : 44406081
25 : 61358726
26 : 137777
27 : 8812689
28 : 8812739
29 : 2459318
30 : 4777982

Fig.2 – Top 30 drug leads for coronavirus identified with their PubChem ID generated by the program

```
os.system(
    'vina --config coordinatesandsetups1p9u.txt --receptor 1p9u.pdbqt --ligand {} --out {} --log {}'.format(
    str(i)+'.pdbqt', str(i)+'.1p9u.pdbqt', str(i)+'.1p9u.log')
)
print("Docking to 1p9u protein completed. Files saved as {} as docked complex and {} as a logfile".format(
    str(i)+'.1p9u.pdbqt',str(i)+'.1p9u.log'))




print(url)


input("Press any key to continue In Silico modelling for next lead compound")
```

```
https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/0012759/SDF
Press any key to continue In Silico modelling for next lead compound
Preparing ligand 2459318 for docking
Running docking procedure for 6w9c
Docking to 6x9c protein completed. Files saved as 2459318.6w9c.pdbqt as docked complex and 2459318.6w9c.log as a
logfile
Running docking procedure for 1p9u
Docking to 1p9u protein completed. Files saved as 2459318.1p9u.pdbqt as docked complex and 2459318.1p9u.log as a
logfile
https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/2459318/SDF
Press any key to continue In Silico modelling for next lead compound
Preparing ligand 4777982 for docking
Running docking procedure for 6w9c
Docking to 6x9c protein completed. Files saved as 4777982.6w9c.pdbqt as docked complex and 4777982.6w9c.log as a
logfile
Running docking procedure for 1p9u
Docking to 1p9u protein completed. Files saved as 4777982.1p9u.pdbqt as docked complex and 4777982.1p9u.log as a
logfile
https://pubchem.ncbi.nlm.nih.gov/rest/pug/compound/cid/4777982/SDF
Press any key to continue In Silico modelling for next lead compound
```

Fig.3 – Automated *In Silico* modelling

```
for image in pngs:
    print(image.get_attribute("href"))
    output_image = requests.get(image.get_attribute("href"))
    open(
        os.getcwd()+'/{}'.format(cplx+'.png'), 'wb'
    ).write(output_image.content)
    print("Image saved as {}".format(cplx+'.png'))

for pysession in pymolsessions:
    print(pysession.get_attribute("href"))
    pse = requests.get(pysession.get_attribute("href"))
    open(
        os.getcwd()+'/{}'.format(cplx+'.pse'), 'wb'
    ).write(pse.content)
    print("Pymol sessions saved as {}".format(cplx+'.pse'))

restart_plip = plip.find_element_by_xpath('/html/body/div[1]/div[2]/div/p[3]/a').click()
time.sleep(5)
except:
    print("No interactions found for {} or damaged structure".format(cplx))
    try:
        restart_plip = plip.find_element_by_xpath('/html/body/div[1]/div[2]/div/p[3]/a').click()
        time.sleep(5)
    except:
        restart_plip = plip.find_element_by_xpath('/html/body/div[1]/div[2]/div[2]/p/a').click()
```

```
Connection successful
Processing file 1p9u-35470824.cplx.pdb
https://projects.biotec.tu-dresden.de/plip-web/plip/download/e07e1d04-421e-422e-a877-1117986d13a3?filePath=output
s%2F1P9U_35470824_PROTEIN_UNL_d_1.png
Image saved as 1p9u-35470824.cplx.pdb.png
https://projects.biotec.tu-dresden.de/plip-web/plip/download/e07e1d04-421e-422e-a877-1117986d13a3?filePath=output
s%2F1P9U_35470824_PROTEIN_UNL_d_1.pse
Pymol sessions saved as 1p9u-35470824.cplx.pdb.pse
```

Fig.4 – Automated Protein-Ligand profiling of interaction between lead compounds and coronaviral drug targets

The crystal structure of the coronavirus main protease which is identified as a coronavirus drug target was downloaded from the RCSB-PDB database [21] with PDB ID : 1P9U. The crystal structure of papain like protease which is identified as a coronavirus drug target was downloaded from the RCSB-PDB database [21] with PDB ID : 6W9C. The ligand and the protein files were prepared for AutoDock process using AutoDockTools(ADT) scripts and the protein drug target files are to be kept in the working folder of the user and can be downloaded from the GitHub repository. The structure of the drug lead compounds generated by the program were programmatically downloaded from PubChem and programmatically prepared for molecular docking using AutoDockTools ligand preparation scripts. The virtual screening using AutoDock Vina [22,23] was initiated programmatically through the program and the interaction between the drug targets proteins and lead drug compounds also automatically profiled and the visualized image of the protein-ligand interaction is saved in the working folder of the user by the program [24,25]. The program also allows the user to control the number of compounds that must proceed for *In Silico* analysis in the following way as shown in Fig.3 thus allowing the user to control the program as required by the research needs. The program stores the results of the *In Silico* modelling in the working folder of the user and accesses them programmatically as shown in Fig.4 to produce protein-ligand interaction profiling images and stores the results in the working folder of the user.

Results and Discussion

The python program was run in Python 3 environment with the dependency packages mentioned in the methodology section. The program prints out the PubChem CIDs of the top 30 compounds identified as coronavirus drug leads by the program and is shown in Fig.2. This is done by the program through automated programmatic data mining of PubChem database to collect data required to perform a machine learning based AutoQSAR algorithm on the dataset to generate the drug leads for coronavirus. The drug leads generated by the program for the coronavirus are useful to screen PubChem database which is over a billion compounds and the generated drug leads are useful to further pursue *In Silico, In Vitro and In Vivo* testing and is expected to save computational and experimental testing costs for the pharmaceutical industry.

The drug leads generated by the program were required to satisfy the Lipinski's criteria of drug likeness as compounds that satisfy the Lipinski's criteria are likely to be orally active drug in humans. The structure of the compounds identified as drug leads were

programmatically downloaded from PubChem by the program and they were fed as programmatic ligand input files after ligand preparation via ADT scripts to a *In Silico* modelling package used widely known as AutoDock-Vina. Therefore the study of the interaction of the drug lead compounds and the coronavirus drug target proteins papain like protease (PDB ID : 6W9C) and coronavirus main protease (PDB ID : 1P9U) was automated through the programmatic inputs given to AutoDock-Vina in the program. The results of the virtual screening for the top 30 drug lead compounds are given in Table 1 and Table 2. It can be found that the compounds generated as drug leads by the program interacted favourably with the two drug targets of coronavirus thus making a case for the usefulness of the tool.

Table 1 – Interaction of drug lead compounds and main protease of coronavirus

Ligand	Target protein	AutoDock Vina score (Kcal/mol)
CID 19655808		-6.8
CID 1604503		-6.9
CID 45053517		-6.3
CID 6510409		-6.7
CID 5442666.	1P9U	-7.0
CID 2131970		-6.0
CID 3095861		-6.1
CID 2131982		-7.1
CID 2926675		-6.2
CID 2131972		-6.0
CID 1719488		-6.6
CID 102341659		-7.1
CID 135411479		-6.7
CID 16194028		-6.6
CID 2928447		-6.0
CID 5736415		-6.4
CID 135470824		-6.8
CID 5470478		-5.3
CID 5442667		-6.5
CID 5849846		-7.7
CID 143349191		-7.9
CID 73155023		-7.1
CID 121513876		-7.5
CID 127043644		-7.4
CID 127043645		-7.9
CID 127045227		-7.4
CID 5336255		-6.7
CID 127045222		-6.6
CID 143349197		-7.5
CID 5470475		-7.3

Table 2 – Interaction of drug lead compounds and papain like protease

Ligand	Target protein	AutoDock Vina score (Kcal/mol)
CID 19655808		-7.5
CID 1604503		-7.1
CID 45053517		-6.8
CID 6510409		-7.4
CID 5442666	6W9C	-7.2
CID 2131970		-7.4
CID 3095861		-7.4
CID 2131982		-6.7
CID 2926675		-6.6
CID 2131972		-6.6
CID 1719488		-6.1
CID 102341659		-7.7
CID 135411479		-8.1
CID 16194028		-7.1
CID 2928447		-7.0
CID 5736415		-7.2
CID 135470824		-6.9
CID 5470478		-6.8
CID 5442667		-6.7
CID 5849846		-7.4
CID 143349191		-7.6
CID 73155023		-8.0
CID 121513876		-7.3
CID 127043644		-7.9
CID 127043645		-7.4
CID 127045227		-7.3
CID 5336255		-7.6
CID 127045222		-7.1
CID 143349197		-8.3
CID 5470475		-7.8

Select images of interaction of the compounds generated as drug leads by the program and the coronaviral drug targets, main protease (PDB ID : 1P9U) and that papain like protease (PDB ID : 6W9C) is shown in Fig 5a, 5b, 5c, 5d & 5e and Fig 6a, 6b, 6c, 6d & 6e respectively.
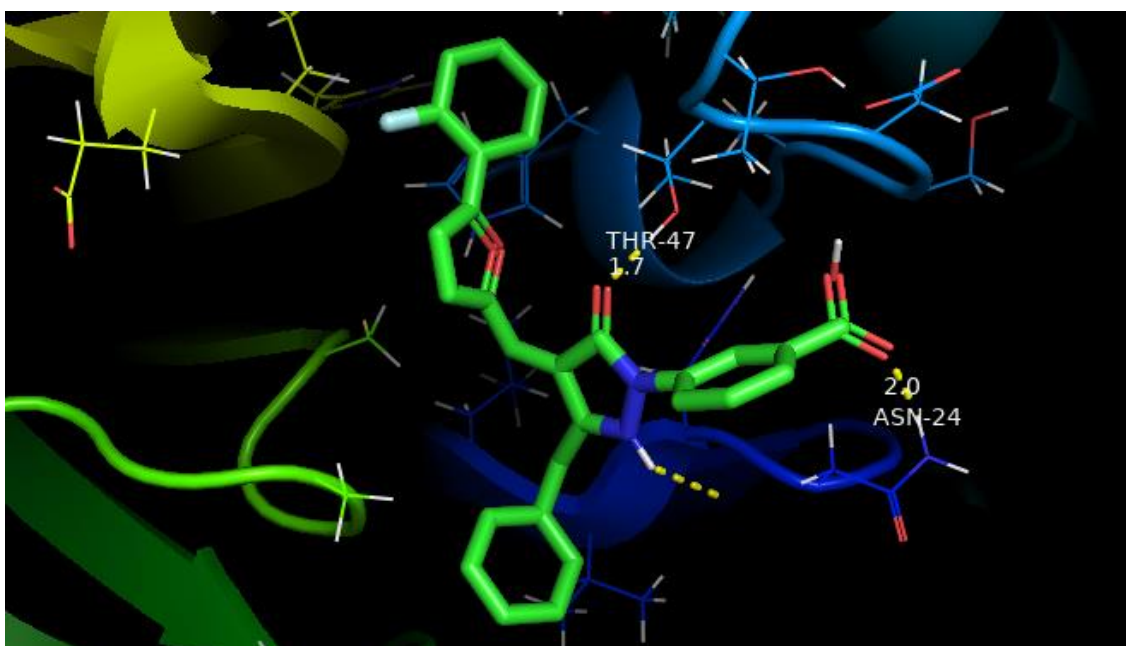
Fig.5a – Interaction of coronavirus main protease and compound identified with PubChem CID 4334
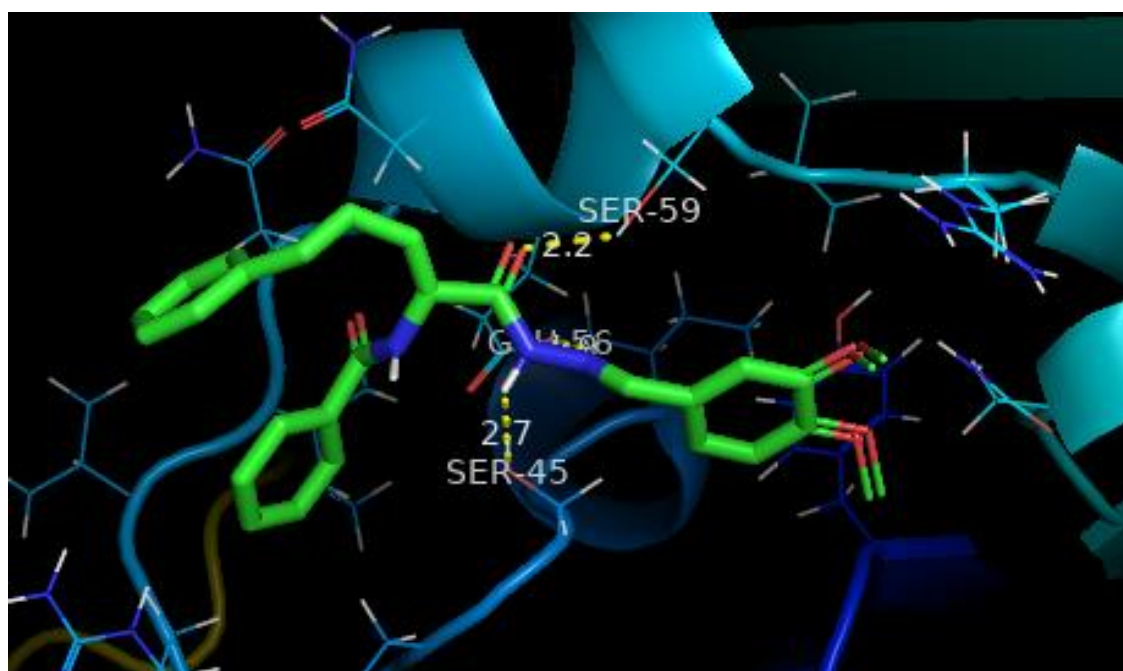

Fig. 5b - Interaction of coronavirus main protease and compound identified with PubChem CID 5336255
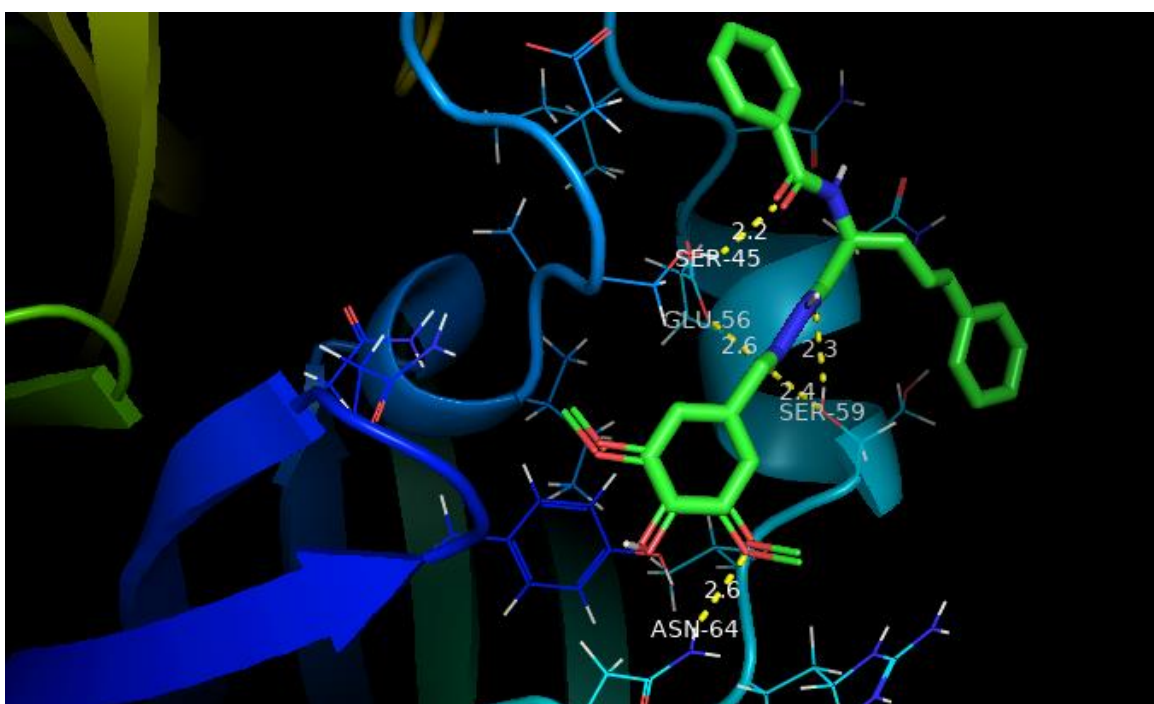
Fig. 5c – Interaction of coronavirus main protease and compound identified with PubChem CID 35470824
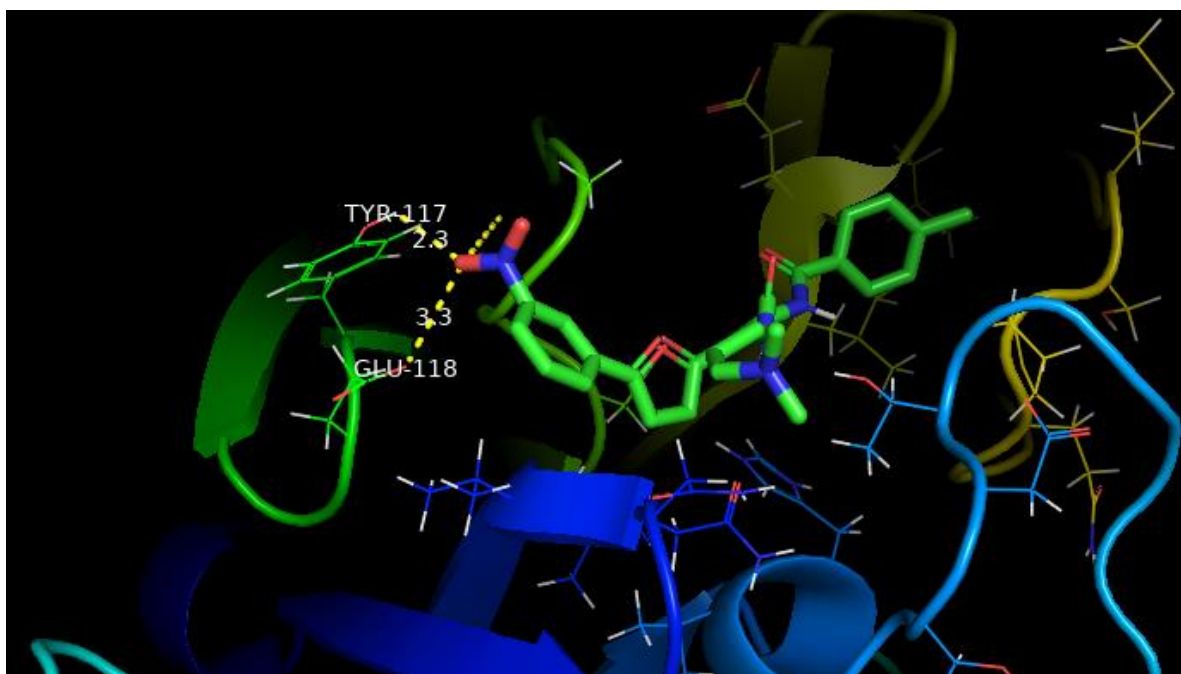

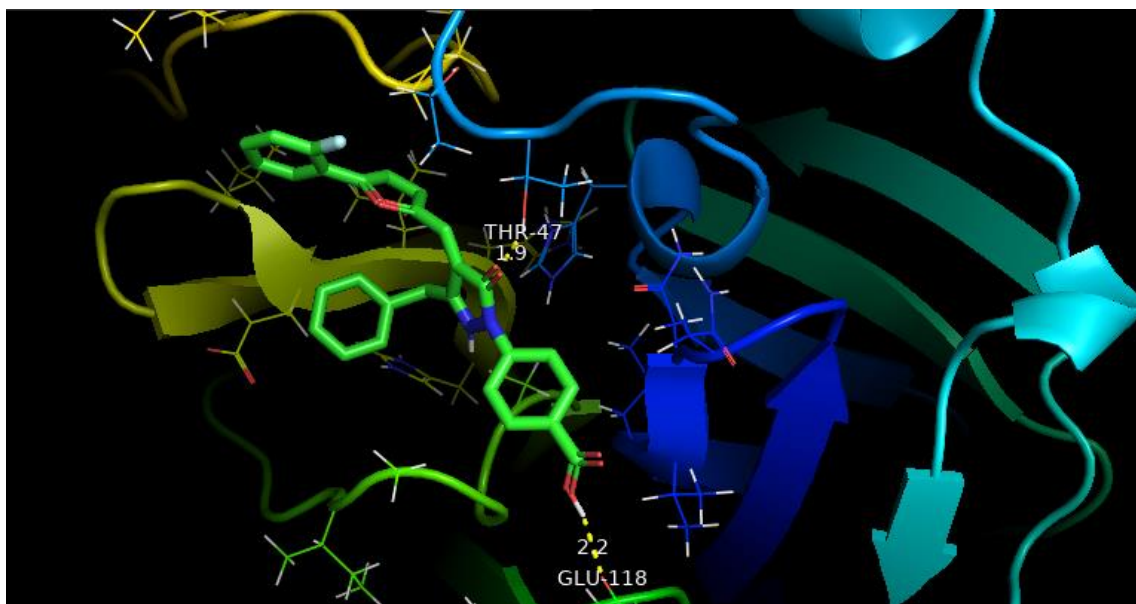Fig. 5d – Interaction of coronavirus main protease and compound identified with PubChem CID 309586

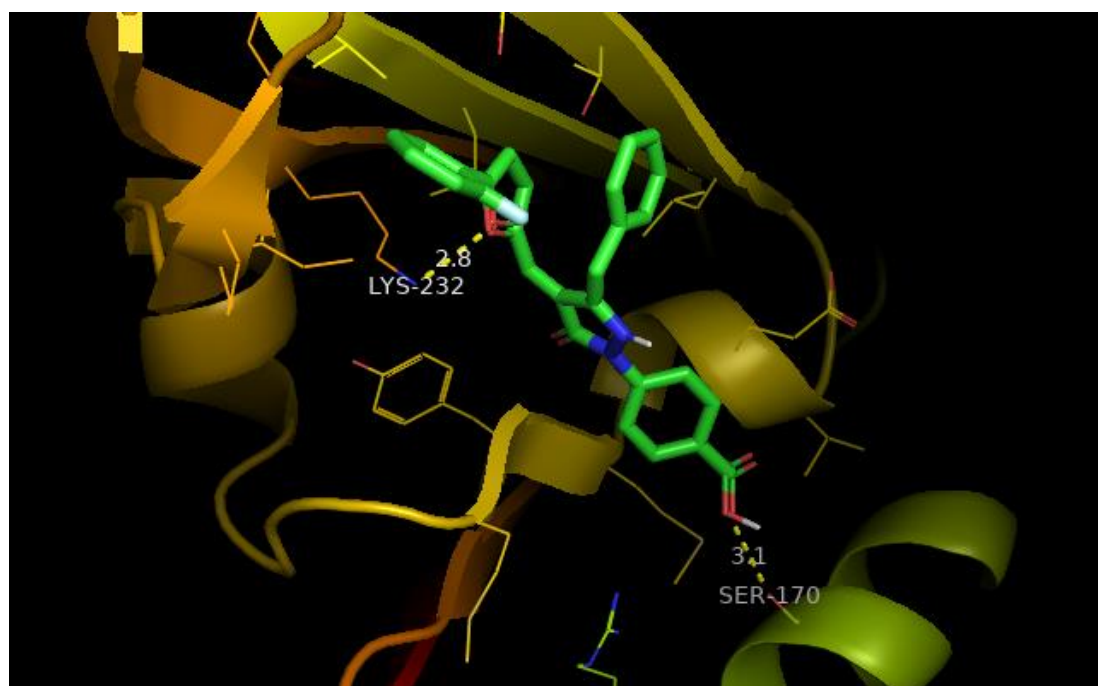Fig. 5e – Interaction of coronavirus main protease and compound identified with PubChem CID 309586

Fig. 6a – Interaction of coronavirus drug target papain like protease and compound identified with PubChem CID 12151387
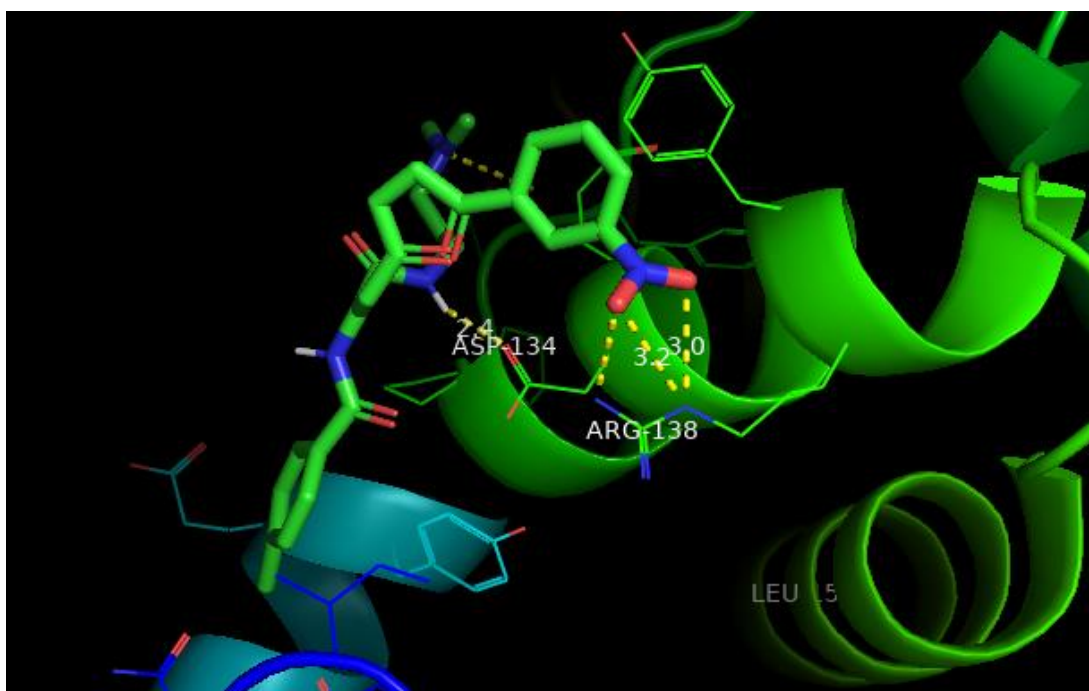
Fig. 6b - Interaction of coronavirus drug target papain like protease and compound identified with PubChem CID 2131970
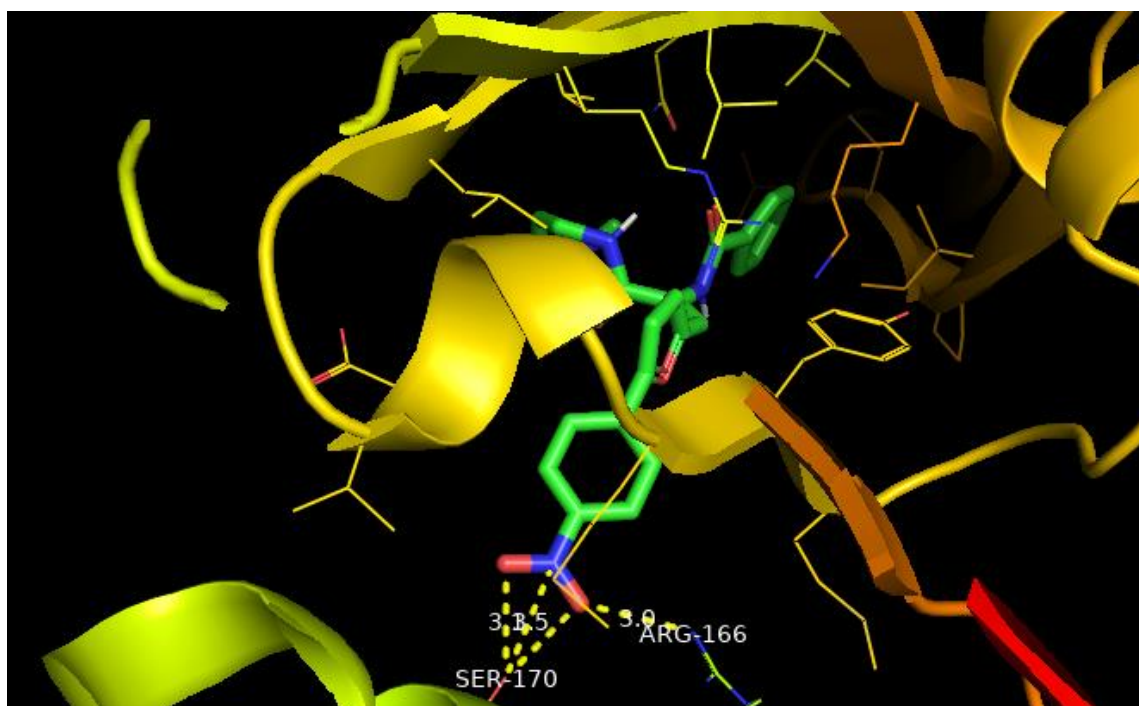

Fig. 6c - Interaction of coronavirus drug target papain like protease and compound identified with PubChem CID 1719488
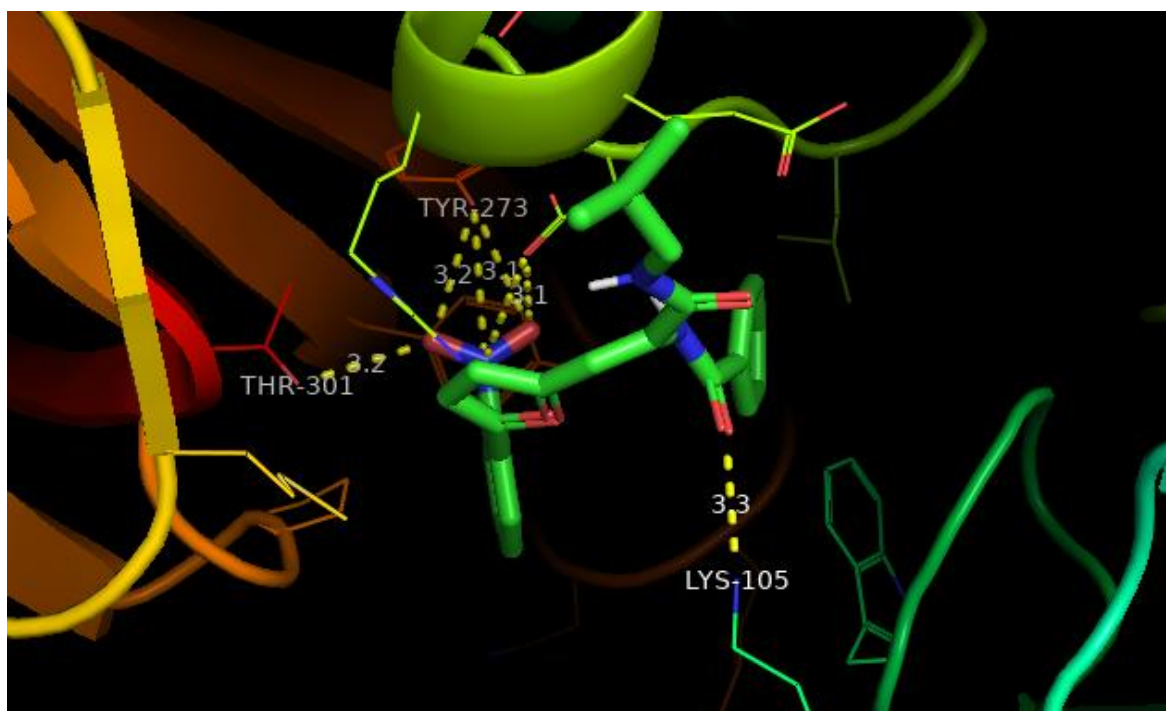
Fig. 6d - Interaction of coronavirus drug target papain like protease and compound identified with PubChem CID 45053517

Fig. 6e - Interaction of coronavirus drug target papain like protease and compound identified with PubChem CID 135470824

Therefore we see the usefulness of the tool in reducing the complexity of virtual screening for drugs against coronavirus achieved through programmatic automation of data mining of PubChem to collect data required to implement a machine learning based AutoQSAR algorithm for automatic drug lead generation. The program requires that the drug leads generated by the program were required to satisfy the Lipinski's rule of five or the drug-likeness criteria. Further, the program automates the In Silico modelling of the interaction of the compounds generated as drug leads and the drug targets of coronavirus and stores the results in the working folder of the user. Thus, the program helps achieve complete automation in identifying drugs against coronavirus which further has to be examined for drug potential through experimental testing such as *In Vitro* and *In Vivo* testing.

Conclusion and future scope

Thus the presented work is an attempt to automate the workflow of drug discovery for coronavirus by a python program automate the process of data mining PubChem database to collect data required to perform a machine learning based AutoQSAR algorithm through which drug leads for coronavirus are generated. The data acquisition from PubChem was carried out through python web scrapping techniques. The workflow of a machine learning based AutoQSAR involves feature learning and descriptor selection, QSAR modelling, validation and prediction. The drug leads generated by the program are required to satisfy the Lipinski's drug likeness criteria. Drug leads generated by the program are fed as programmatic inputs to an In Silico modelling package to computer model the interaction of the compounds generated as drug leads and two coronavirus drug targets identified with their PDB ID : 6W9C and 1P9U. The results are stored in the working folder of the user. The program also generates protein-ligand interaction profiling and stores the visualized images in the working folder of the user. Thus our programmatic tool ushers in the new age automatic ease in drug identification for coronavirus through a fully automated QSAR and an automated In Silico modelling of the drug leads generated by the autoQSAR algorithm. While the compounds identified through the automated workflow must be test experimentally by the experimental research community for their drug potential against coronavirus, there is still a lot of scope to make the automation algorithm more self-aware of technical nuances which will help increase its accuracy in drug identification which we bring to the attention of the computational research community for their scholarly attention and efforts on the same.

References

1. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Bryant, S. H. (2015). PubChem Substance and Compound databases. Nucleic Acids Research, 44(D1). doi:10.1093/nar/gkv951

2. 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

3. Swain, M. (2014). PubChemPy: A way to interact with PubChem in Python.

4. Bhardwaj, V. (2014). Quantitative structure–activity relationship (QSAR) studies as strategic approach in drug discovery. Medicinal Chemistry Research, 23(12), 4991-5007. doi:10.1007/s00044-014-1072-3

5. Eriksson, L., & Johansson, E. (1996). Multivariate design and modeling in QSAR. Chemometrics and Intelligent Laboratory Systems, 34(1), 1-19. doi:10.1016/0169-7439(96)00023-8.

6. Bajot, F. (2009). The Use of Qsar and Computational Methods in Drug Design. *Challenges and Advances in Computational Chemistry and Physics Recent Advances in QSAR Studies,* 261-282. doi:10.1007/978-1-4020-9783-6_9

7. Sippl, W. (2009). 3D-QSAR – Applications, Recent Advances, and Limitations. Challenges and Advances in Computational Chemistry and Physics Recent Advances in QSAR Studies, 103-125. doi:10.1007/978-1-4020-9783-6_4.

8. Mitchell, R. (2018). *Web Scraping with Python: Collecting More Data from the Modern Web*. O'Reilly Media, Incorporated.

9. Vanden Broucke, S., & Baesens, B. (2018). *Practical Web scraping for data science* (pp. 3-5). New York, NY: Apress.

10. Ton, A. T., Gentile, F., Hsing, M., Ban, F., & Cherkasov, A. (2020). Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. *Molecular informatics*.

11. Dixon, S. L., Duan, J., Smith, E., Von Bargen, C. D., Sherman, W., & Repasky, M. P. (2016). AutoQSAR: an automated machine learning tool for best-practice quantitative structure–activity relationship modeling. *Future medicinal chemistry*, *8*(15), 1825-1839.

12. Kim, S., & Cho, K. H. (2019). PyQSAR: A Fast QSAR Modeling Platform Using Machine Learning and Jupyter Notebook. *Bulletin of the Korean Chemical Society*, *40*(1), 39-44.

13. Lipinski, C. A. (2004). Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, *1*(4), 337-341.

14. Kumar, V., Jung, Y. S., & Liang, P. H. (2013). Anti-SARS coronavirus agents: a patent review (2008–present). *Expert opinion on therapeutic patents*, *23*(10), 1337-1348.

15. Nascimento Junior, J. A. C., Santos, A. M., Quintans-Júnior, L. J., Walker, C. I. B., Borges, L. P., & Serafini, M. R. (2020). SARS, MERS and SARS-CoV-2 (COVID-19) treatment: a patent review. *Expert Opinion on Therapeutic Patents*, (just-accepted).

16. Fischer, A., Sellner, M., Neranjan, S., Smieško, M., & Lill, M. A. (2020). Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. *International Journal of Molecular Sciences*, *21*(10), 3626.

17. Rifaioglu, A., Sinoplu, E., Atalay, V., Martin, M., Cetin-Atalay, R., & Dogan, T. (2020). DEEPScreen: High Performance Drug-Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations. *Chemical Science*.

18. Gentile, F., Agrawal, V., Hsing, M., Ton, A. T., Ban, F., Norinder, U., ... & Cherkasov, A. (2020). Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Central Science*.

19. Liao, Z., You, R., Huang, X., Yao, X., Huang, T., & Zhu, S. (2019, November). DeepDock: Enhancing Ligand-protein Interaction Prediction by a Combination of Ligand and Structure Information. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 311-317). IEEE.

20. Rodgers, S. L., Davis, A. M., Tomkinson, N. P., & van de Waterbeemd, H. (2011). Predictivity of simulated ADME AutoQSAR models over time. *Molecular informatics*, *30*(2-3), 256-266.

21. Deshpande, N., Addess, K. J., Bluhm, W. F., Merino-Ott, J. C., Townsend-Merino, W., Zhang, Q., ... & Kramer Green, R. (2005). The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic acids research, 33(suppl_1), D233-D237.

22. Jaghoori, M. M., Bleijlevens, B., & Olabarriaga, S. D. (2016). 1001 Ways to run AutoDock Vina for virtual screening. *Journal of computer-aided molecular design*, *30*(3), 237-249.

23. Forli, S., Huey, R., Pique, M. E., Sanner, M. F., Goodsell, D. S., & Olson, A. J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. Nature protocols, 11(5), 905.

24. Tibaut, T., Borišek, J., Novič, M., & Turk, D. (2016). Comparison of in silico tools for binding site prediction applied for structure-based design of autolysin inhibitors. *SAR and QSAR in Environmental Research*, *27*(7), 573-587.

25. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F., & Schroeder, M. (2015). PLIP: fully automated protein–ligand interaction profiler. *Nucleic acids research*, *43*(W1), W443-W447.