QSARStudy On P38 Mitogen Activated Protein Kinase - Inhibition For Bronchial Asthma

1Rayan Ahmed
1UG Scholar
1Karunya Institute of Technology and Sciences

Abstract - **Asthma is a disease of the airways caused by chronic inflammation, elevated serum IgE levels, airway hyperresponsiveness and, frequently, airway remodelling. T-helper type 2 cells together with mast cells, B cells, and eosinophils as well as inflammatory cytokines and chemokines are proposed to play a critical role in the initiation and perpetuation of allergic asthma. therefore, the inhibition of p38 MAPK is a lucrative therapeutic strategy for the treatment of asthma. The p38 MAPK are activated using variety of stimuli. A large dataset of 2,186 compounds with reported IC50 values against MAPK p38α was obtained from ChEMBL database and induced in quantitative structure-activity relationship study so as to gain insights on their origin of bioactivity. MAPK p38 α were described by a set of 12 fingerprint descriptors and predictive models were constructed from regressor comparison best model where selected and further tuned to give better performance. It was observed that several 12 fingerprint descriptors afforded good performance were in R2 0.32 – 0.63, Q2CV 0.24 – 0.53 and Q2Ext 0.26 – 0.55 for the training set. It is anticipated that our proposed QSAR model may become a useful high-throughput tool for identifying novel inhibitors against p38 MAPKs. Conclusion: Twelve sets of fingerprint descriptors were used for constructing QSAR models and their performances were comparatively evaluated. It was observed that several fingerprint descriptors afforded good performance for the constructed models indicating that they could capture the feature space of p38 MAPKs inhibitors. By taking advantage of the built-in feature importance estimator and SHAP Library plots, the following important features that are for MAPKs inhibition were identified: SubFPC301(1,5-Tautomerizable), SubFPC295(C ONSbond), SubFPC274(Aromatic ring) and SubFPC181(HeteroNnonbasic). Results from molecular docking also support the aforementioned findings from the QSAR models in which the aromatic, and heteroaromatic rings were preferable moieties for interacting with the hydrophobic pocket of p38 MAPKα. It is anticipated that the knowledge gained from this study could be used as general guidelines for the design of novel p38 MAPKα inhibitors.**

keywords - **QSAR study, Docking study, p38 MAPK, Inhibitor, Asthma, Data Science**

INTRODUCTION:

Asthma is a heterogeneous disease, usually characterized by chronic airway inflammation. It is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable expiratory airflow limitation is a common, chronic respiratory disease affecting 1 – 18% of the population in different countries, it can be characterized by variable symptoms of wheeze, shortness of breath, chest tightness and /or cough and by variable expiratory airflow limitation.(Global Initiative for Asthma, 2021). Rising of allergic respiratory diseases among different group of people has been observed. Causes of these diseases can be explained by the presence of biologic aeroallergens such as Pollen and HDM which are able to stimulate the sensitization and symptoms of these diseases. Sensitization to different clinical symptoms and severity of the diseases; rhinitis, asthma and rhinitis with/without asthma may be caused by different types of aeroallergens (A Valero; S Quirce; I Dávila; J Delgado, 2017). Asthma is one of the threatening respiratory diseases affecting both children and adult and usually characterized by chronic airway inflammation. It is defined by the history of respiratory symptoms such as wheeze, shortness of breath, chest tightness and cough that vary over time and in intensity, together with variable airflow obstruction (AL Durham; G Caramori; KF Chung; IM Adcock, 2016). The MAPK family includes the p38 kinases, which consist of highly conserved proline-directed serine-threonine protein kinases that are activated in response to inflammatory signals. Of the four isoforms, p38α is the most abundant in inflammatory cells and has been the most studied through mainly the availability of small molecule inhibitors. The p38 substrates include transcription factors; other protein kinases, which in turn phosphorylate transcription factors; cytoskeletal proteins and translational components; and other enzymes. Both asthma and COPD are characterized by chronic airflow obstruction, airway and lung remodelling, and chronic inflammation. p38 is involved in the inflammatory responses induced by cigarette smoke exposure, endotoxin, and oxidative stress through activation and release of proinflammatory cytokines/chemokines, posttranslational regulation of these genes, and activation of inflammatory cell migration. Inhibition of p38 MAPK prevented allergen-induced pulmonary eosinophilia, mucus hypersecretion, and airway hyperresponsiveness, effects that may partly result from p38 activation on eosinophil apoptosis and on airway smooth muscle cell production of cytokines/chemokines. In addition, p38 regulates the augmented contractile response induced by oxidative stress. (Kian Fan Chung, 2011) There are four isoforms of p38 MAPK, p38α, p38ß, p38γ and p38δ. α and ß isoforms seem to be more relevant to asthmatic inflammation, as in the human lung cDNA libraries, their levels are higher than those of γ and δ isoforms, and are associated with activation of more downstream targets (Jiang, et al., 1997) p38 protein, reorganization mainly occurs inside the

hydrophobic docking groove. (Mayor et al., 2007) The Database for Bioactivity data of compounds "ChEMBL is an Open Data database containing binding, functional and ADMET information for a large number of drug-like bioactive compounds. These data are manually abstracted from the primary published literature on a regular basis, then further curated and standardized to maximize their quality and utility across a wide range of chemical biology and drug-discovery research problems. Currently, the database contains 17 million bioactivity measurements for more than 2.1 million compounds and 14K protein targets. Access is available through a web-based interface, data downloads and web services" (Davies et al., 2015). The reason for using Computational approaches have often been used to complement experimental studies for several reasons, which among others include: (i) handle and manage large volumes of biological and chemical information, (ii) model biomolecular phenomenon that may be impossible by experimental means and (iii) making sense of data by uncovering hidden patterns and trends. In the context of drug discovery efforts, *in silico* approaches can be used to not only help identify and prioritize classes of compounds to screen but it can also help reduce the number of compounds to be tested. Quantitative structure–activity relationship is ligand-based approach in computational drug design for correlating the molecular features of a chemical library with their respective bioactivity. QSAR has been instrumental in shedding light on the molecular basis of bioactivity of interest by learning from past bioactivity data while also being amenable to extrapolating on the bioactivity of new compounds that are foreign to the trained data set. (Suvannang et al., 2018). QSAR is a paradigm that enables the prediction of biological activities for compounds of interest as a function of their descriptors through the use of statistical or machine learning methods (Nantasenamat et al., 2009). Aside from the ability to predict the activity, QSAR models have been instrumental in enabling understanding on the origin of these biological activities by means of interpreting the descriptors used in building such models. Herein, the first large-scale QSAR investigation for predicting p38 MAPK inhibition, which to the best of our knowledge the largest collection of 2,186 non-redundant compounds. QSAR models were built using interpretable learning methods (e.g., lightgbm) and descriptors (i.e., molecular fingerprints) to unravel the underlying MAPK inhibitory activity, which was performed in accordance with guidelines of the OECD. Molecular docking was also performed on an Exclusion set of compounds selected from active MAPK inhibitors. The ligand and structure-based approach employed in this study is anticipated to be useful in the design and development of p38 MAPKs inhibitors.

AIM

Analysis on the Biological Activity Data Retrieved from ChEMBL Database for the potential Inhibitors – "p38 Mitogen Activated Protein Kinase-α" responsible asthmatic inflammatory network.

OBJECTIVES

- ✓ Recognize the Targets and active molecules for asthmatics inhibition
- ✓ Preform QSAR Studies Using Data Science by creating Novel Dataset
- ✓ Bioactivity Dataset Preparation
- ✓ Exploratory Data Analysis
- ✓ Descriptor Dataset Preparation
- ✓ Comparisons of Machine Learning Models
- ✓ Model Interpretability
- ✓ Molecular Docking
- ✓ Interpretation of Binding Modality

REVIEW OF LITERATURE

Insite on Asthma

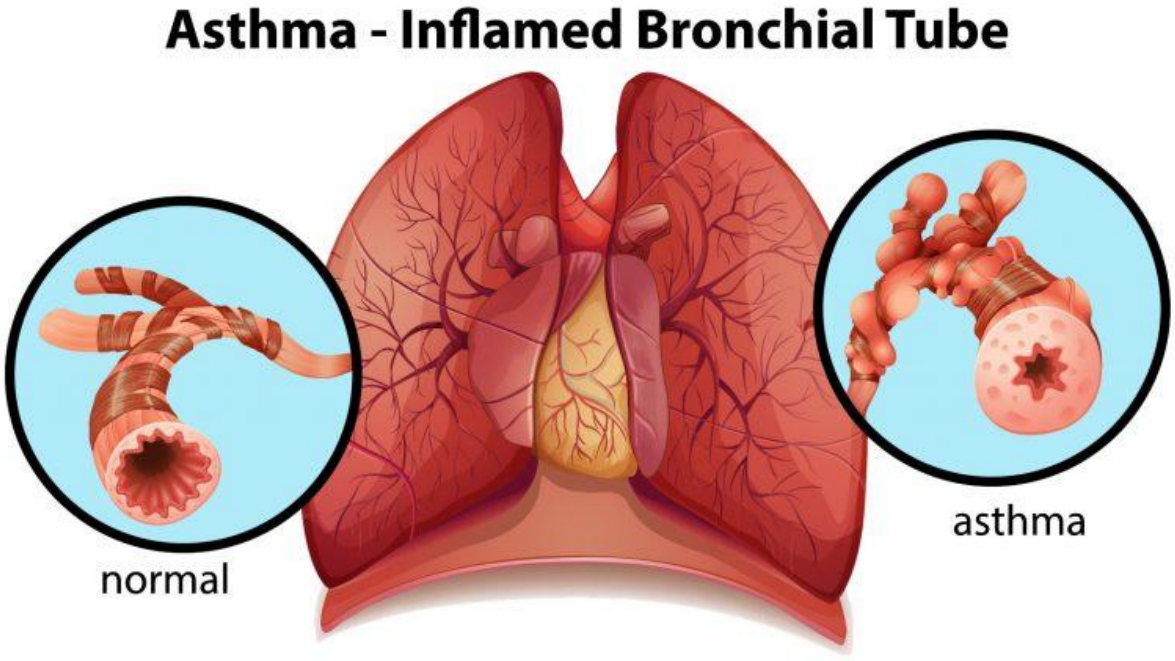


Fig.1 Asthma Airway [22]

Asthma has been described as a medical entity since the time of Aretaeus, the Cappadocian, in approximately 100 A.D. the constellation of physical findings and signs that we currently recognize as asthma dates from the work of John Floyer in 1698. Floyer defined asthma as "laborious respiration with lifting of the shoulders and wheezing." He understood that asthma was intermittent and episodic and that the treatment of asthma needs to consist of rescue and controller therapy, termed by him as treatment "both in fit and out of it."

By 1900, it was well established that certain forms of asthma could be brought on by exposure to environmental allergens. In Stedman's *Twentieth Century Practice*, published in 1896, Sir Thomas Granger Stewart and George Alexander Gibson wrote about asthma

The treatment of asthma involves the treatment of the patient during fits and between the fits. The general indications are: To allay the spasm during the paroxysm; To find out and remove the exciting cause., To treat complications and sequelae and to improve the general health.

Thus, more than 100 years ago, the general approach to asthma treatment was then as it is now: acute rescue treatment, controller treatment, and prevention of long-term complications. (Chu EK & Drazen JM., 2005)

Introduction into clinical use of the first modern inhaler for the management of asthma — the pressurised metered-dose inhaler (pMDI). The pMDI was initially used for the administration of the non-selective beta-agonists adrenaline and isoprenaline. However, the epidemic of asthma deaths which occurred in the 1960s led to these drugs being superseded by the selective short-acting beta-agonist salbutamol, and the first inhaled corticosteroid (ICS) beclomethasone. At the same time, sodium cromoglycate was introduced, to be administered via the first dry-powder inhaler — the Spinhaler — but owing to its relatively weak anti-inflammatory action its use is now very limited. Over the last 10 years, the long-acting beta-agonists (LABAs) have become an important add-on therapy for the management of asthma, and they are now often used with ICS in a single ICS/LABA combination inhaler.(Crompton., 2006).

Insite on p38 MAPKs:

Among the various members of the MAPK family, p38 MAPK subgroup is the most involved in airway and lung inflammation underlying asthma and COPD. In particular, several environmental agents including aeroallergens, cigarette smoke, airborne pollutants, viral and bacterial pathogens activate the p38α isoform which in turn up-regulates the expression of multiple proinflammatory cytokines and chemokines, as well as the production of some fibrogenic factors. Therefore, p38 MAPK-induced bronchial inflammation and remodelling significantly contribute to the development, persistence and amplification of airflow limitation, which is the hallmark of asthma and COPD Within such a pathobiologic framework, the main triggers of asthma and COPD are aeroallergens, cigarette smoke, airborne pollutants, and viral/bacterial infections. Inside the respiratory tract, all these environmental agents are able to activate the p38 subfamily of mitogen-activated protein kinases (Pelaia et al., 2021). p38 kinases have two domains: a 135 amino acid N-terminal domain and a 225 amino acid C-terminal domain. The main secondary structure of the N-terminal domain is β-sheets, while the C-terminal domain has a α-helical structure. The catalytic site is located in the region linking the two domains. The phosphorylation lip of p38 consists of 13 residues, Leu-171–Val-183, and the protein is activated by phosphorylation of a single threonine (Thr-180) and a single tyrosine residue (Tyr-182) in the lip. Moreover, in *Drosophila* p38 MAPK, phosphorylation of tyrosine-186 was detected exclusively in the nucleus following osmotic stress. p38 isoforms show various three-dimensional structures with differences in the orientation of the N- and C-terminal domains, resulting in different sized ATP-binding pockets (Yang et al., 2014). Pharmaceutical companies and researchers have worked hard to develop novel, safe, and specific p38 inhibitors. Based on the importance of p38α in inflammation, people have focused on inhibitors for this isoform rather than the other isoforms. ML3403, a SB203580 analogue, represses the expression of TNF-α, IL-6, and IL-8. It can bind to both active and inactive forms of p38α kinase, which may reduce asthma-induced airway inflammation and remodelling. MAP kinase plays a very important role in diseases, such as asthma, osteoarthritis, rheumatoid arthritis RA and chronic inflammatory autoimmune diseases. Therefore, the inhibition of MAP kinase would potentially prevent the underlying pathophysiology in the inflammatory diseases which makes it an attractive target for drug discovery. (Jatavath,Sivan Lingala & Manga, 2011)

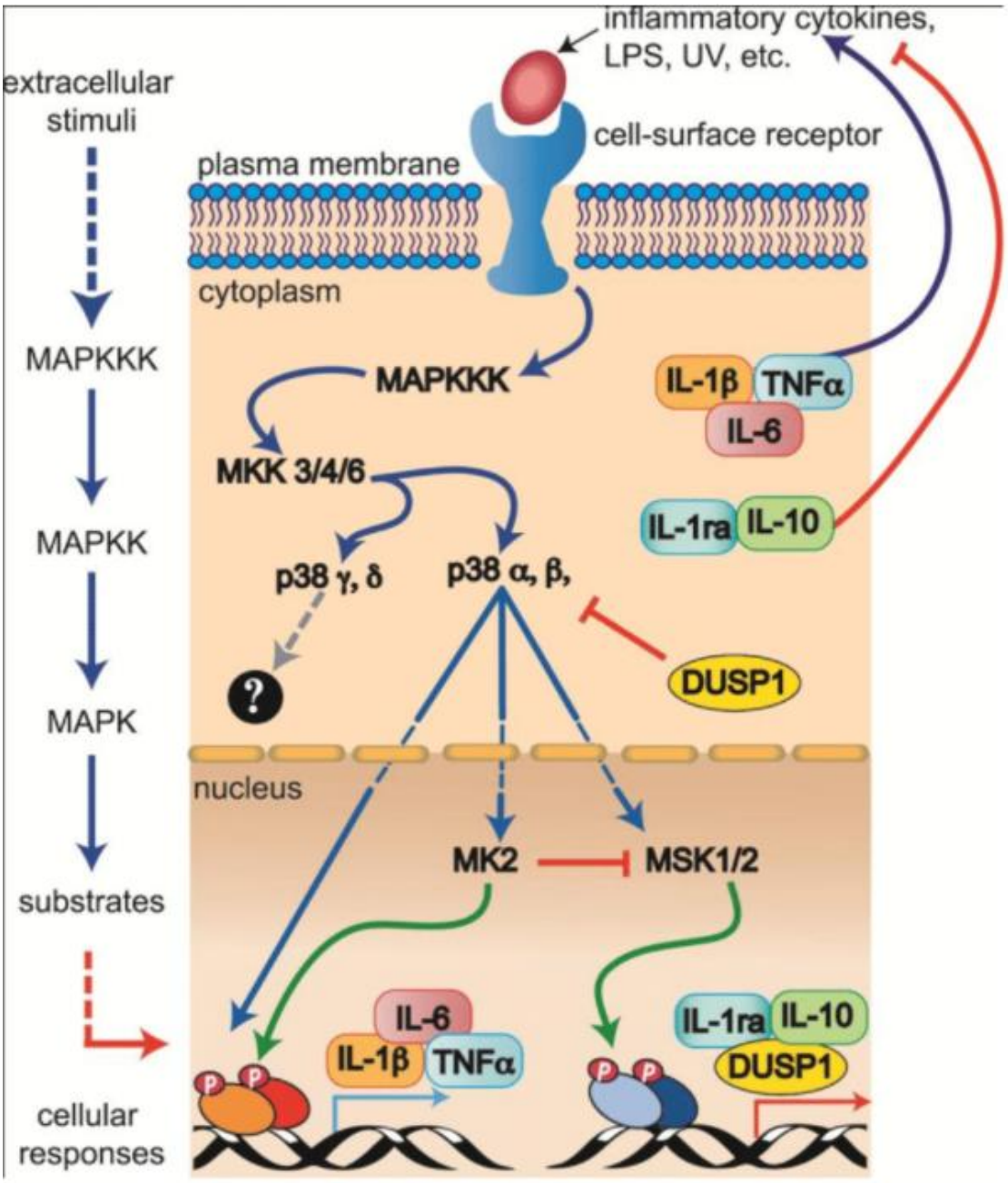


Fig.2[16] Pathway of p38 MAPKs

There is increase number of studies going on for the therapeutic target for MAPK as there are numerus research work being done on. The number of clinical trials has increased to achieve potent drug for treating the inflammatory diseases in link with p38 MAPKα. With more than 20 compounds entering human trials, none have progressed beyond Phase II to the best of our knowledge (Millan, 2011)

TYPES OF ASTHMA:

Adult-onset asthma can be further broken down into the subtypes of:

- Allergic asthma
- Non-allergic asthma
- Cough-variant asthma
- Glucocorticoid-Resistant Asthma

Allergic Asthma Between 50% and 80% of all asthma cases are allergic asthma, which is also called extrinsic asthma.3 In people with seasonal allergies (hay fever), this type may be called seasonal asthma. With allergic asthma, your immune system responds inappropriately to allergens, which triggers both allergy symptoms and asthma symptoms (airway obstruction, wheezing).

Non-Allergic Asthma Between 10% and 33% of all people with asthma have non-allergic asthma, which is sometimes called intrinsic asthma.[3] It usually develops later in life than allergic asthma. Some research suggests non-allergic asthma is more severe than other forms as measured by the Global Initiative for Asthma (GINA) score.

Cough-Variant Asthma While a mucousy cough may accompany the usual symptoms of asthma, some people's only asthma symptom is a dry cough. This is called cough variant asthma. The cough may remain your sole symptom or you may go on to develop other symptoms, especially if your asthma isn't properly treated.
Glucocorticoid-Resistant Asthma While glucocorticoids are one of the most potent anti-inflammatory drugs available and are normally very effective asthma treatments, they don't work well in between 5% and 10% of people with asthma. When that happens, the patient is said to have glucocorticoid-resistant or steroid-resistant asthma, another subtype of AOA Viral-induced asthma is caused by a respiratory tract infection, such as the common cold, flu, or COVID-19. An estimated 50% of acute asthma attacks have a viral trigger. (Pat Bass, MD, verywellhealth.com, 2021)

CAUSES OF ASTHMA:

What Does Asthma Feel Like?

Asthma is marked by inflammation of the bronchial tubes, with extra sticky secretions inside the tubes. People with asthma have symptoms when the airways tighten, inflame, or fill with mucus.

There are three major signs of asthma:

- **Airway blockage.** When you breathe as usual, the bands of muscle around your airways are relaxed, and air moves freely. But when you have asthma, the muscles tighten. It's harder for air to pass through.
- **Inflammation**. Asthma causes red, swollen bronchial tubes in your lungs. This inflammation can damage your lungs. Treating this is key to managing asthma in the long run.
- **Airway irritability.** People with asthma have sensitive airways that tend to overreact and narrow when they come into contact with even slight triggers.
  (Pichardo, 2019)

The fundamental causes of asthma are not completely understood. The strongest risk factors for developing asthma are a combination of genetic predisposition with environmental exposure to inhaled substances and particles that may provoke allergic reactions or irritate the airways, such as:

- indoor allergens (for example house dust mites in bedding, carpets and stuffed furniture, pollution and pet dander)
- outdoor allergens (such as pollens and moulds)
- tobacco smoke
- chemical irritants in the workplace
- air pollution

Other triggers can include cold air, extreme emotional arousal such as anger or fear, and physical exercise. Even certain medications can trigger asthma: aspirin and other non-steroid anti-inflammatory drugs, and beta-blockers Urbanization has been associated with an increase in asthma. But the exact nature of this relationship is unclear. (WHO, 2020)

TREATMENTS OF ASTHMA:

Patients with severe asthma respond less well to corticosteroids than those with non-severe asthma. Increased p38 mitogen-activated protein kinase (MAPK) activation in alveolar macrophages (AMs) from severe asthma patients has been associated with a reduced inhibition of cytokine release by dexamethasone. (Bhavsar, Khorasani, Hew.,2010)

Long-term asthma control medications, generally taken daily, are the cornerstone of asthma treatment. These medications keep asthma under control on a day-to-day basis and make it less likely you'll have an asthma attack. Types of long-term control medications include:

- **Inhaled corticosteroids.** These medications include fluticasone propionate (Flovent HFA, Flovent Diskus, Xhance), budesonide (Pulmicort Flexhaler, Pulmicort Respules, Rhinocort), ciclesonide (Alvesco), beclomethasone (Qvar Redihaler), mometasone (Asmanex HFA, Asmanex Twisthaler) and fluticasone furoate (Arnuity Ellipta).

You may need to use these medications for several days to weeks before they reach their maximum benefit. Unlike oral corticosteroids, inhaled corticosteroids have a relatively low risk of serious side effects.

- **Leukotriene modifiers.** These oral medications — including montelukast (Singulair), zafirlukast (Accolate) and zileuton (Zyflo) — help relieve asthma symptoms.

Montelukast has been linked to psychological reactions, such as agitation, aggression, hallucinations, depression and suicidal thinking. Seek medical advice right away if you experience any of these reactions.

- **Combination inhalers.** These medications — such as fluticasone-salmeterol (Advair HFA, Airduo Digihaler, others), budesonide-formoterol (Symbicort), formoterol-mometasone (Dulera) and fluticasone furoate-vilanterol (Breo Ellipta) — contain a long-acting beta agonist along with a corticosteroid.
- **Theophylline.** Theophylline (Theo-24, Elixophyllin, Theochron) is a daily pill that helps keep the airways open by relaxing the muscles around the airways. It's not used as often as other asthma medications and requires regular blood tests.(MAYO, 2020)

STUDIES

Application of 3D-QSAR on a Series of Potent P38-MAP Kinase Inhibitors

One of the most applied methods in drug industry for development of new drugs is 3D-QSAR methodology. As p38-MAPK plays a crucial role in regulating the production of such proinflammatory cytokines as tumor necrosis factor-α and interleukin-1, emerging as an attractive target for new anti-inflammatory agents, we used a 3D-QSAR based method of CoMFA on a

series of 52 potent p38-MAP kinase inhibitors with IC 50 ranging from 3.2 to 10,000 nM. An alignment rule for the compounds was defined using Distill in SYBYL 7.3. The best model was validated using a data set obtained by dividing the data set into a training set and test set using hierarchical clustering, based on the CoMFA fields and biological activities (pIC 50). The best predictions were obtained with a CoMFA region-focusing model (R 2 ncv = 0.952, q 2 = 0.678, R 2 Pred = 0.627). The statistical parameters from the model indicate that the data are well fitted and has high predictive ability. Moreover, the resulting 3D CoMFA contour maps provide useful guidance for designing highly active inhibitors.
In this study, by using the alignment scheme generated from Distill, a predictive CoMFA region focusing model was developed and was used to predict the pIC50 activity of a set of MAP-kinase inhibitors. The QSAR model gave good statistical results in terms of q2 and R2 values, and has been validated using a test set, obtained from the hierarchical clustering. From this study, it is possible to predict the ligand activities of newly designed MAP-kinase inhibitors, and design better inhibitors.
(Safavi & Ghasemi, 2013)

Pharmacophore generation, atom-based 3D-QSAR, docking, and virtual screening studies of p38-α mitogen activated protein kinase inhibitors: pyridopyridazin-6-ones pt.2
p38-α mitogen-activated protein kinase (MAPK) is a serine/threonine kinase activated by environmental stimuli, like stress, and by various proinflammatory cytokines, such as tumor necrosis factor-α and interleukin-1β. Excessive production of tumor necrosis factor-α and interleukin-1β may lead to various diseases, such as rheumatoid arthritis, psoriasis, and inflammatory bowel disease. Hence, inhibition of p38-α MAPK could be a novel approach for the development of new anti-inflammatory agents. In this study, a combination of pharmacophore generation, an atom-based three-dimensional quantitative structure-activity relationship (3D-QSAR), molecular docking, and virtual screening was performed for a series of pyridopyridazin-6-ones exhibiting p38-α MAPK inhibition activity. A five-point pharmacophore (AAAHR), ie, three hydrogen bond acceptors (AAA), one hydrophobic (H) group, and one aromatic ring (R) was obtained. A statistically significant 3D-QSAR model was obtained using this pharmacophore hypothesis with a good correlation coefficient (R^2=0.91) and a high Fisher ratio (F=90.3) for the training set of 47 compounds. The predictive power of the model generated was found to be significant, and was confirmed by the high value of the cross-validated correlation coefficient (q^2=0.80) and Pearson's R (0.90) for the test set of 16 compounds. Further, the docking study revealed the binding orientations of the active ligand on the amino acid residues Valine 30 (Val30), Glycine 31 (Gly31), Lysine 53 (Lys53), Leucine 75 (Leu75), Aspartic acid 88 (Asp88), Methionine 109 (Met109) of p38-α MAPK at the active site. The results of this ligand-based pharmacophore hypothesis and atom-based 3D-QSAR provide detailed structural insights and highlight the important binding features between pyridopyridazin-6-ones and p38-α MAPK. These findings may provide useful guidelines for rational design of compounds with better p38-α MAPK activity. Finally, four hits were obtained after virtual screening, and docking studies that have a good docking score, and predictive and fitness value may be further modified using the QSAR model to obtain the most potent p38-α MAPK inhibitor. Hence, the results of an atom-based 3D-QSAR model and docking studies help in designing analogs with better activity prior to synthesis which may be potent p38-α MAPK inhibitors. (S Bhansali & VM Kulkarni., 2014)

3D-QSAR models to predict anti-cancer activity on a series of protein P38 MAP kinase inhibitors
Protein kinases are essential components of various signaling pathways and represent attractive targets for therapeutic interventions. Kinase inhibitors are currently used to treat malignant tumors, as well as autoimmune diseases, due to their involvement in immune cell signaling. In this study, three-dimensional quantitative structure–activity relationship (3D-QSAR) analyses, including Multiple Linear Regression , Partial Least Squares , Multiple Non-Linear Regression , Artificial Neural Network and cross-validation analyses, were performed on a set of P38 MAP kinases as anti-cancer agents. This method, which is based on molecular modeling (molecular mechanics, Hartree-Fock), was used to determine the structural parameters, electronic properties, and energy associated with the molecules we examined. MLR, PLS, and MNLR analyses were performed on 46 protein P38 MAP kinase analogs to determine the relationships between molecular descriptors and the anti-cancer properties of the P38 MAP kinase analogs. The MLR model was validated by the external validation and standardization approach. The ANN, given the descriptors obtained from the MLR, exhibited a correlation coefficient close to 0.94. The predicted model was confirmed by two methods, leave-one-out cross-validation and scrambling (or Y-randomization). We observed a high correlation between predicted and experimental activity, thereby both validating and demonstrating the high quality of the QSAR model that we described. A QSAR was performed on 46 molecules derived from P38 MAP kinases. A QSAR model was established using the Multiple Linear Regression , Partial Least Squares , Multiple Non-Linear Regression and Artificial Neural Network paradigms. The resulting model can be used to predict the anticancer activity of P38 MAP kinases. (Hadaji, Bourass, Oummou & Bouachrine., 2016)

The Conclusion from the literature review would be to incorporate new advancement in ways for doing the QSAR the use for new Machine Learning libraries to find the best predicting model for the dataset you have. Calculation and prediction of the bioactivity data can also be done using molecular fingerprint which are the building blocks for each molecule. Molecular fingerprints are a robust descriptor type with immense utility in cheminformatics and computer-aided drug design owing to its information-rich description on the structural details of investigated compounds. The advantage of these descriptors is that they can be rapidly generated in a high-throughput fashion while also affording robust performance and interpretability.

Machine Learning Libraries

PyCaret

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It is an end-to-end machine learning and model management tool that speeds up the experiment cycle exponentially and makes you more productive.

In comparison with the other open-source machine learning libraries, PyCaret is an alternate low-code library that can be used to replace hundreds of lines of code with few words only. This makes experiments exponentially fast and efficient. (Ali Moez, 2020)

Molecular Fingerprint

PaDEL-Descriptor is a software for calculating molecular descriptors and fingerprints. The software currently calculates 797 descriptors (663 1D, 2D descriptors, and 134 3D descriptors) and 12 types of fingerprints. These descriptors and fingerprints are calculated mainly using The Chemistry Development Kit. Some additional descriptors and fingerprints were added, which include atom type electro topological state descriptors, McGowan volume, molecular linear free energy relation descriptors, ring counts, count of chemical substructures identified by Laggner, and binary fingerprints and count of chemical substructures identified by Klekota and Roth. (Yap, 2011)

MATERIALS AND METHODS

A Summary of workflow followed in the study is provide in Fig.3 which induces QSAR model at a large-scale for the predicting and analysing the p38 MAPKα Inhibition

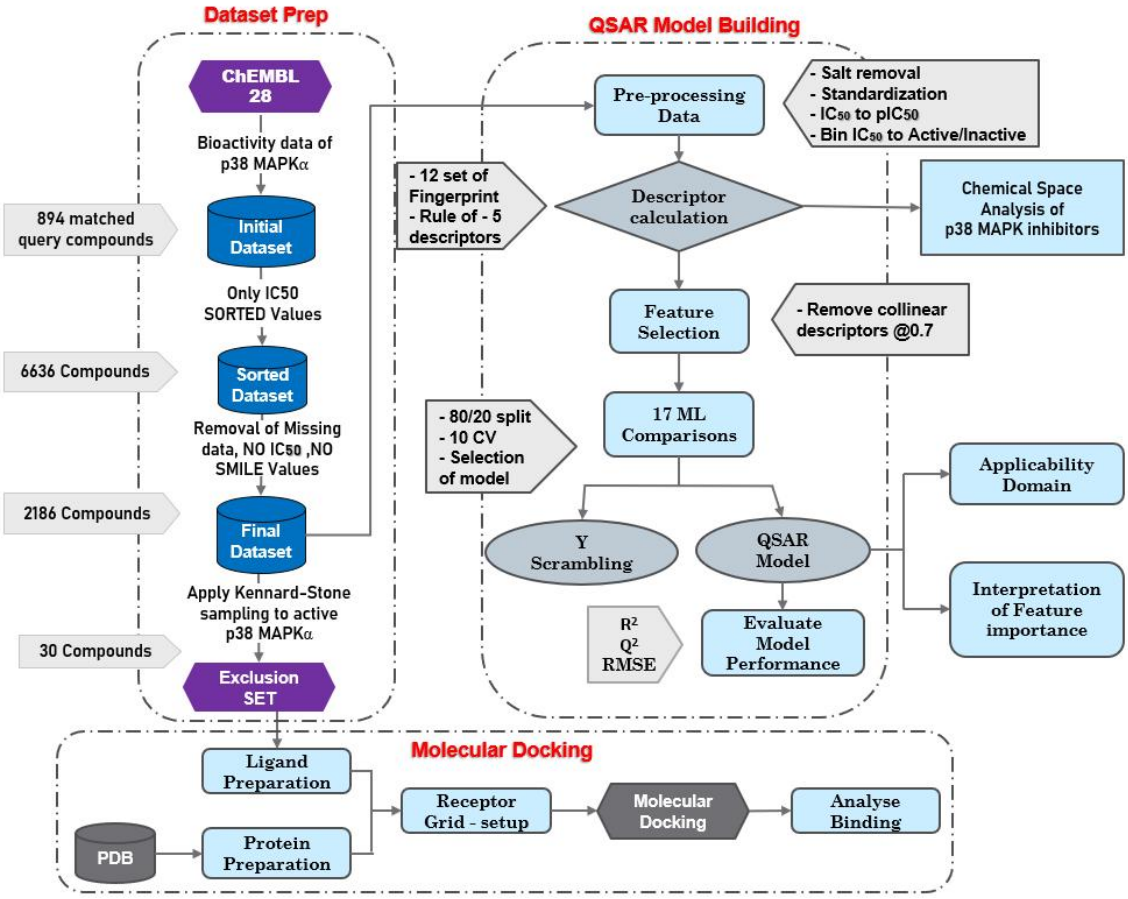


Fig.3 Workflow QSAR modelling and Molecular docking

Data Set

A data set of inhibitors against human p38 MAPKα (Target ID: CHEMBL260) were compiled from the ChEMBL 28 database that is comprised of 849 matched query. The initial data set was curated from several bioactivity measurement units including IC_{50}, K_i, MIC, EC_{50}, etc. IC_{50} was selected for further investigation as they constituted the largest subset with 6,636 compounds. A closer look revealed that compounds had no reported IC_{50} values or had lesser/greater than signs which were subjected to removal thereby. Only data having unit nM as the bioactivity were selected, which produced 2,595 compounds. Moreover, some compounds were found to have no SMILES notation associated with it and were thus removed. A final data set comprising of 2,186 compounds was obtained. The removal of non-redundant SMILES and Bioactivity Value left us with 2,186 unique compounds.

The Kennard -Stone algorithm (Kennard & Stone, 1969) was applied at the active MAPKs inhibitors (i.e., IC_{50} <1 µM) which in return gave 30 compounds which then referred as Exclusion set as these will be further as ligand for molecular docking

The 2,186 was Bin into Compounds having values of less than 1000 nM will be considered to be **active** while those greater than 10,000 nM will be considered to be **inactive**. As for those values in between 1,000 and 10,000 nM will be referred to as **intermediate**.

Description of Inhibitors

p38 MAPKα inhibitors were encoded by a vector of fingerprint descriptors accounting for its molecular constituents. Prior to calculating descriptors, salts were removed and tautomers were standardized using the built-in function of the PaDEL-Descriptor software (Yap, 2011)

No.	Fingerprint Descriptors	Total No.	Description	References
1	CDK	1024	Fingerprint of length 1024 and search depth of 8	Steinbeck et al. (2003)
2	CDK extended	1024	Extends the fingerprint with additional bits describing ring features	Steinbeck et al. (2003)
3	CDK graph only	1024	A special version that considers only the connectivity and not bond order	Steinbeck et al. (2003)
4	E-state	79	Electrotopological state atom types	Hall & Kier (1995)
5	MACCS	166	Binary representation of chemical features defined by MACCS keys	Durant et al. (2002)
6	PubChem	881	Binary representation of substructures defined by PubChem	NCBI (2009)
7	Substructure	307	Presence of SMARTS patterns for functional groups	Laggner (2005)
8	Substructure count	307	Count of SMARTS patterns for functional groups	Laggner (2005)
9	Klekota–Roth	4860	Presence of chemical substructures	Klekota & Roth (2008)
10	Klekota–Roth count	4860	Count of chemical substructures	Klekota & Roth (2008)
11	2D atom pairs	780	Presence of atom pairs at various topological distances	Carhart, Smith & Venkataraghavan (1985)
12	2D atom pairs count	780	Count of atom pairs at various topological distances	Carhart, Smith & Venkataraghavan (1985)

Table.1 Summary of fingerprints

Moreover, Lipinski's rule-of-five comprising of molecular weight (MW), logarithm of the octanol/water partition coefficient (ALogP), number of hydrogen bond donor (nHBDon) and number of hydrogen bond acceptor (nHBAcc) was also calculated.

Feature Selection

Collinearity is a condition where descriptor pairs are known to have intercorrelation, which not only add complexity to the model but could potentially give rise to bias. Pearson's correlation coefficient greater than the threshold of 0.7 was filtered out using the *findCorrelation* function from the R package *caret* to obtain a smaller subset of descriptor (Kuhn, 2008)

Machine Learning Model Comparison

The PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It has module "compare_model" which compare the provided dataset with all model available and displays the best performing model. (Ali Moez, 2020)

Type of Regressor
Light Gradient Boosting Machine
Random Forest Regressor
K Neighbors Regressor
Gradient Boosting Regressor
Bayesian Ridge
Ridge Regression
AdaBoost Regressor
Linear Regression
Huber Regressor
Extra Trees Regressor
Orthogonal Matching Pursuit
Decision Tree Regressor
Lasso Regression

Elastic Net
Lasso Least Angle Regression
Passive Aggressive Regressor
Least Angle Regression

Table.2 Working Regressors

Supervised learning is to learn a model from labeled training data which can be used to make prediction about unseen or future data (James et al., 2013). This study constructs regression models, which affords the prediction of the continuous response variable (i.e., pIC_{50}) as a function of predictors (i.e., fingerprint descriptors).

Model Assessment

Model assessment of QSAR model is the evaluation of the model's performance and robustness or validity of the model prior to its usage in predicting and interpreting the biological activities of compounds. Two statistical parameters were used to check the performance of the QSAR models consisting of Pearson's correlation coefficient (r) and root mean squared error (RMSE). R is a correlation that is numbered between +1 and -1. It shows the relationship between the dependent variable and the independent variables. Values close to +1 or -1 shows a strong relationship between both variables. RMSE is a commonly used parameter to assess the relative error of the predictive model. The predictive performance was validated using 10-fold cross-validation, external validation and Y-scrambling test.

10-fold Cross-validation is a technique to evaluate predictive models by dividing the original sample into a training set to train the model, and a test set to evaluate it. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples the average accuracies can be used to truly assess the performance of the predictive model.

Y Scrambling is a method that one can use in order to test whether the predictions made by the model aren't made just by chance the Y-dependent variable (i.e., pIC_{50}) was randomly scrambled and the statistical assessment parameters are recalculated.

Applicability Domain Analysis

The applicability domain (AD) estimates the likelihood of reliable prediction for compounds on the basis of how similar they are to compounds used to build the model. Williams Plot was used in this to study analyze the AD. The leverage value allows one to identify whether new compounds will lie within or outside the domain. Leverage values for all compounds are calculated via adjustment of X to give the hat matrix H:

$$H=X(XTX)-1XT \quad - (1)$$

where X is a two-dimensional matrix comprising of n compounds and m descriptors while X^T is the transpose of X. Meanwhile, the leverage value of the ith compound (h_i) is the ith diagonal element of H:

$$hi=xTi(XTX)-1xi \quad - (2)$$

where x_i is the descriptor row-vector of the ith compound. The warning leverage h^* is calculated by:

$$h*=3(p+1)/n. \quad - (3)$$

Practically, the leverage value along with the William's plot is often used to assess the AD of QSAR models. (Simeon S et al., 2016)

Molecular docking

The co-crystal structure of human p38 MAPK with BIRB 796 (PDB ID: 1KV2) was retrieved from the Protein Data Bank and initially prepared using protein preparation wizard. The Gird box was made using the Receptor Gird Generation at coordinates of X: 33.72, Y: 35.64, Z: 17.4. The Exclusion Set of 30 compounds was used in ligand preparation (Schrödinger, 2018-4) The binding energy (kcal/mol) of MAPK inhibitors were calculated. conformers providing the lowest binding energy were selected for further analysis of the binding mode. The Site moiety was done using SiMMap web server (Bollback, 2006)

Reproducible Research

To afford the reproducibility of this research, the code and data used in the construction of QSAR models and analyses performed herein are provided publicly at (Github).

RESULTS & DISCUSSION

Chemical space of p38 MAPK inhibitors

An exploration of the general chemical space of the investigated data set by means of Lipinski's rule-of-five descriptors Visualization of the chemical space of LogP as a function of MW is see in Fig. 4, as to investigate the chemical space of MAPK inhibitors. A dense distribution of inhibitors was observed within the space of MW starting from approximately 300–500 Da and within the space of LogP ranging from approximately 1 to 6.

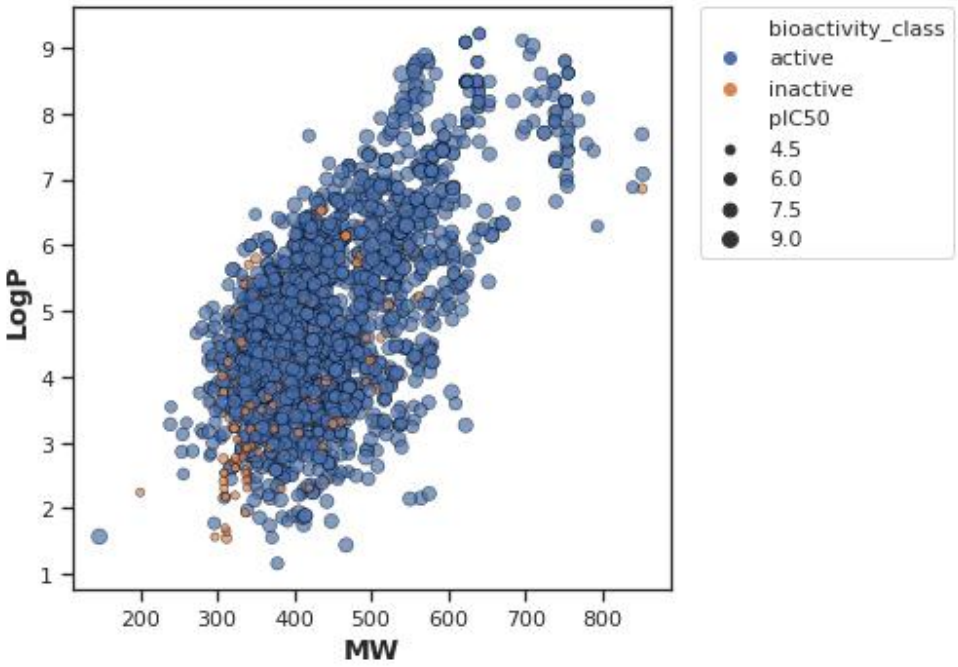


Fig. 4 Chemical space

Overall distribution of data values of Lipinski's descriptors is shown as box plots in Fig. 5 in which the MW, LogP, NumHDonors and NumHAcceptors are shown in Figs. 5A, 5B, 5C and 5D, respectively Of the 4 Lipinski's descriptors (MW, LogP, NumHDonors and NumHAcceptors), only NumHDonors exhibited ***no difference*** between the **actives** and **inactives** while the other 3 descriptors (MW, logP and NumHAcceptors) shows ***statistically significant difference*** between **actives** and **inactives**.

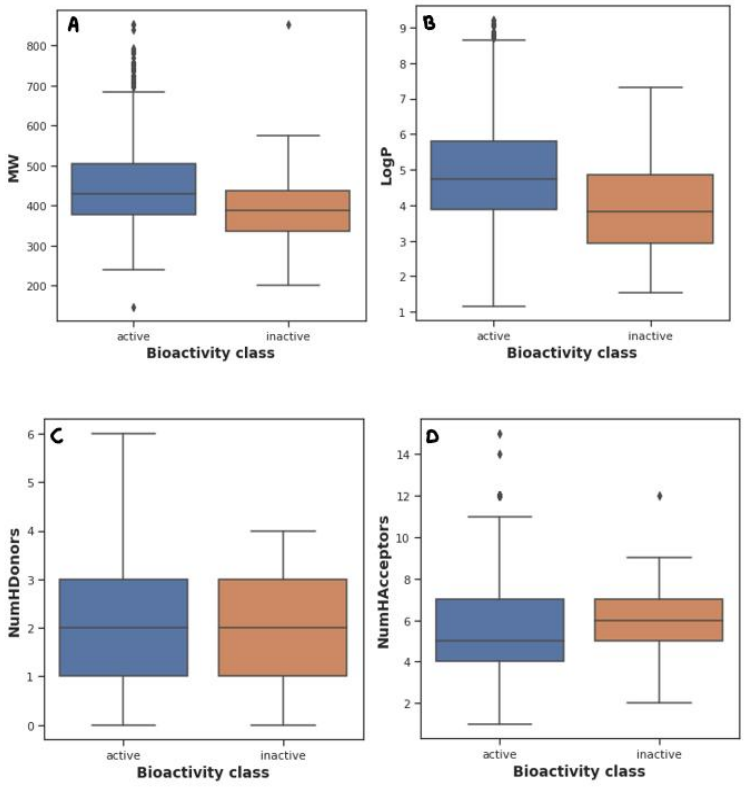
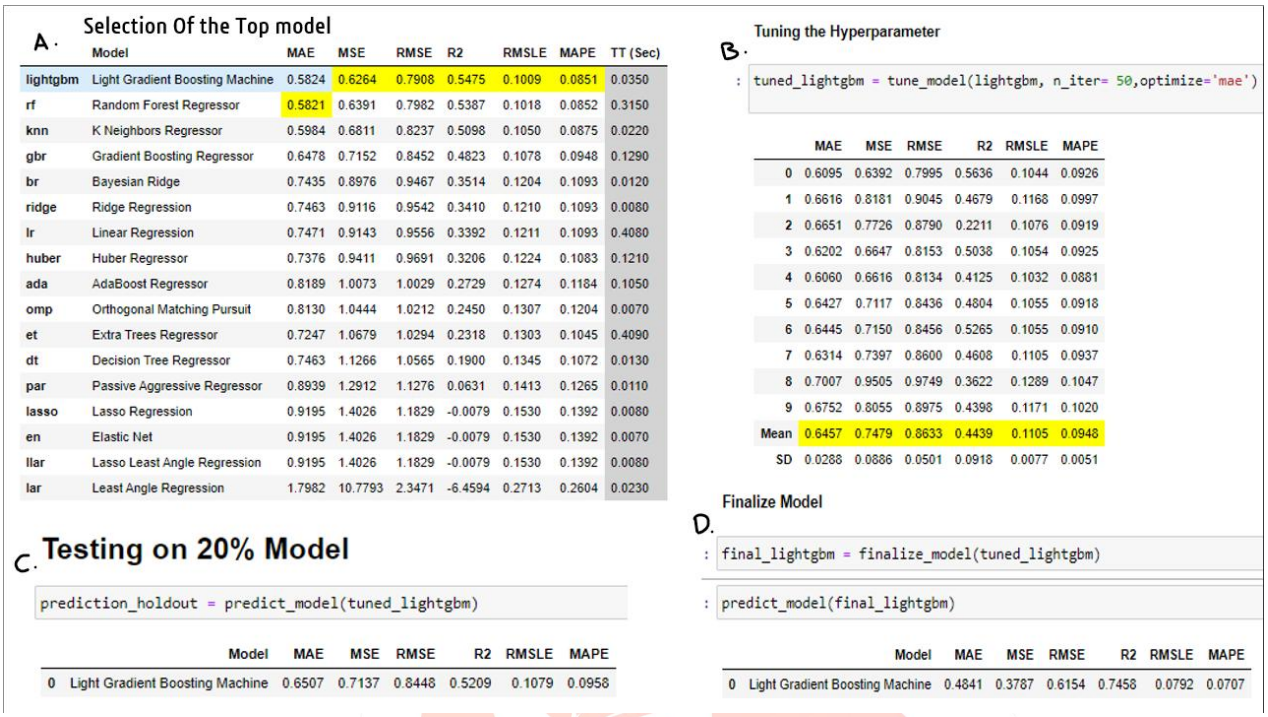


Fig.5 Box plot using Lipinski's rule of five descriptors

ML Comparison

The model comparison was done using the PyCaret compare model package.
the 12 set fingerprint descriptor modelled using the package and found that the selected regressor was "Light Gradient Boosted Machine", or LightGBM for short, is an open-source implementation of gradient boosting designed to be efficient and perhaps more effective than other implementations.As such, LightGBM refers to the open-source project, the software library, and the machine learning algorithm. In this way, it is very similar to the Extreme Gradient Boosting or XGBoost technique.(Guolin Ke et al., 2017)



	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.5824	0.6264	0.7908	0.5475	0.1009	0.0851	0.0350
rf	Random Forest Regressor	0.5821	0.6391	0.7982	0.5387	0.1018	0.0852	0.3150
knn	K Neighbors Regressor	0.5984	0.6811	0.8237	0.5098	0.1050	0.0875	0.0220
gbr	Gradient Boosting Regressor	0.6478	0.7152	0.8452	0.4823	0.1078	0.0948	0.1290
br	Bayesian Ridge	0.7435	0.8976	0.9467	0.3514	0.1204	0.1093	0.0120
ridge	Ridge Regression	0.7463	0.9116	0.9542	0.3410	0.1210	0.1093	0.0080
lr	Linear Regression	0.7471	0.9143	0.9556	0.3392	0.1211	0.1093	0.4080
huber	Huber Regressor	0.7376	0.9411	0.9691	0.3206	0.1224	0.1083	0.1210
ada	AdaBoost Regressor	0.8189	1.0073	1.0029	0.2729	0.1274	0.1184	0.1050
omp	Orthogonal Matching Pursuit	0.8130	1.0444	1.0212	0.2450	0.1307	0.1204	0.0070
et	Extra Trees Regressor	0.7247	1.0679	1.0294	0.2318	0.1303	0.1045	0.4090
dt	Decision Tree Regressor	0.7463	1.1266	1.0565	0.1900	0.1345	0.1072	0.0130
par	Passive Aggressive Regressor	0.8939	1.2912	1.1276	0.0631	0.1413	0.1265	0.0110
lasso	Lasso Regression	0.9195	1.4026	1.1829	-0.0079	0.1530	0.1392	0.0080
en	Elastic Net	0.9195	1.4026	1.1829	-0.0079	0.1530	0.1392	0.0070
llar	Lasso Least Angle Regression	0.9195	1.4026	1.1829	-0.0079	0.1530	0.1392	0.0080
lar	Least Angle Regression	1.7982	10.7793	2.3471	-6.4594	0.2713	0.2604	0.0230

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	0.6095	0.6392	0.7995	0.5636	0.1044	0.0926
1	0.6616	0.8181	0.9045	0.4679	0.1168	0.0997
2	0.6651	0.7726	0.8790	0.2211	0.1076	0.0919
3	0.6202	0.6647	0.8153	0.5038	0.1054	0.0925
4	0.6060	0.6616	0.8134	0.4125	0.1032	0.0881
5	0.6427	0.7117	0.8436	0.4804	0.1055	0.0918
6	0.6445	0.7150	0.8456	0.5265	0.1055	0.0910
7	0.6314	0.7397	0.8600	0.4608	0.1105	0.0937
8	0.7007	0.9505	0.9749	0.3622	0.1289	0.1047
9	0.6752	0.8055	0.8975	0.4398	0.1171	0.1020
Mean	0.6457	0.7479	0.8633	0.4439	0.1105	0.0948
SD	0.0288	0.0886	0.0501	0.0918	0.0077	0.0051

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	0.6507	0.7137	0.8448	0.5209	0.1079	0.0958

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	0.4841	0.3787	0.6154	0.7458	0.0792	0.0707

Fig. 6 Steps for machine learning

Particularly, twelve sets of fingerprint descriptors were benchmarked in order to find the best performing set and model as steps shown in Fig.6 where 6A is compares the available models, 6B Tuning of model using 10-fold CV, 6C Apply the Trained model on the Test set and 6D Finalizing the model by applying the trained model on the whole dataset .Prior to modeling, feature selection was applied to remove collinear descriptors. Each of the twelve models were then built using a data split ratio of 80/20 in which 80% of the data set was used as the internal set and 20% as the external set. The collective result can be seen in Table. 3

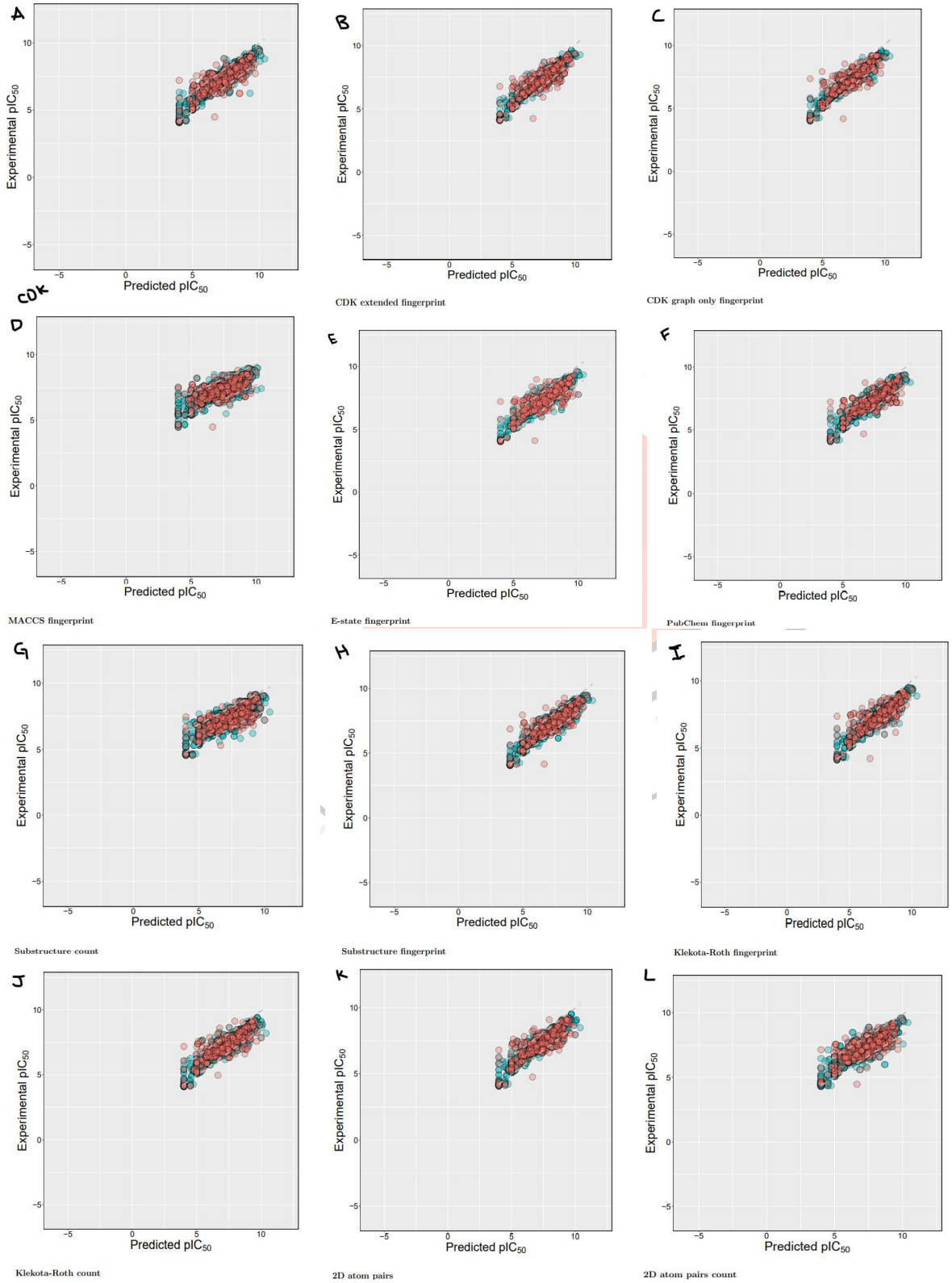
QSAR Model Performance

No.	Fingerprint Descriptors	Total No.	Correlation Remaining	Training set		10–fold CV set		External set		$R2-Q2_{CV}$	$R2-Q2_{Ext}$	Model Trained on Entire Dataset	
				R^2	$RMSE_{Tr}$	$Q2_{CV}$	$RMSE_{CV}$	$Q2_{Ext}$	$RMSE_{Ext}$			R^2	$RMSE_{FT}$
1	CDK	1024	947	0.63	0.86	0.48	0.84	0.55	0.79	0.142	0.081	0.80	0.53
2	CDK extended	1024	937	0.52	0.82	0.53	0.81	0.49	0.80	-0.010	0.026	0.81	0.49
3	CDK graph only	1024	372	0.47	0.85	0.48	0.85	0.48	0.84	-0.005	-0.004	0.71	0.63
4	E-state	79	19	0.32	0.96	0.24	1.02	0.26	1.03	0.085	0.057	0.39	0.94
5	MACCS	166	74	0.46	0.85	0.46	0.84	0.50	0.88	-0.006	-0.044	0.70	0.68
6	PubChem	881	107	0.55	0.79	0.51	0.82	0.55	0.79	0.038	0.002	0.72	0.62
7	Substructure	307	26	0.41	0.89	0.36	0.93	0.31	0.99	0.057	0.100	0.48	0.86
8	Substructure count	307	25	0.46	0.85	0.44	0.85	0.52	0.84	0.016	-0.061	0.75	0.62
9	Klekota–Roth	4860	108	0.48	0.85	0.47	0.85	0.46	0.85	0.002	0.014	0.66	0.68
10	Klekota–Roth count	4860	91	0.51	0.82	0.45	0.87	0.45	0.88	0.057	0.056	0.59	0.76
11	2D atom pairs	780	73	0.44	0.88	0.42	0.89	0.44	0.91	0.021	0.001	0.66	0.70

12	**2D atom pairs count**	**780**	**60**	**0.44**	**0.86**	**0.43**	**0.87**	**0.47**	**0.92**	**0.007**	**-0.030**	**0.67**	**0.72**

Table. 3 Summary QSAR models for predicting pIC_{50}

It can be observed that all twelve models are capable of capturing the inhibitory activity space of p38 MAPKα inhibitors as they provided R^2 and Q^2 (i.e., both 10-fold CV and external sets). It can be seen that models with larger descriptor size, afforded the best performance. The opposite also holds true as the model with the least number of descriptors were also found to perform the worst amongst the other fingerprints. In a nutshell, the model performance in order of decreasing value is as follows: CDK > PubChem > CDK Extended > Klekota–Rota count > Klekota–Roth > CDK graph only > MACCS ≈ Substructure Count > 2D atom pairs ≈ 2D atom pairs count > E-state.



As shown in Fig. 20 for all 12 set of fingerprints, scatter plots of experimental versus predicted pIC_{50} methods as assessed via 10-fold cross-validation and external set. The best performing model is not necessarily the best choice considering the fact that the descriptor size for the best models were quite high and is consequently prone to overfitting. It was found that the substructure count provided reasonably good predictive performance can be seen in Fig.7. To interpret 7A is the residual plot for the Training set, 7B Error plot, 7C Validation Curve Plot and 7D Learning cruve. To further check the reliability and validity of the selected model, Y-scrambling test was performed for 100 iterations seen in Fig.8

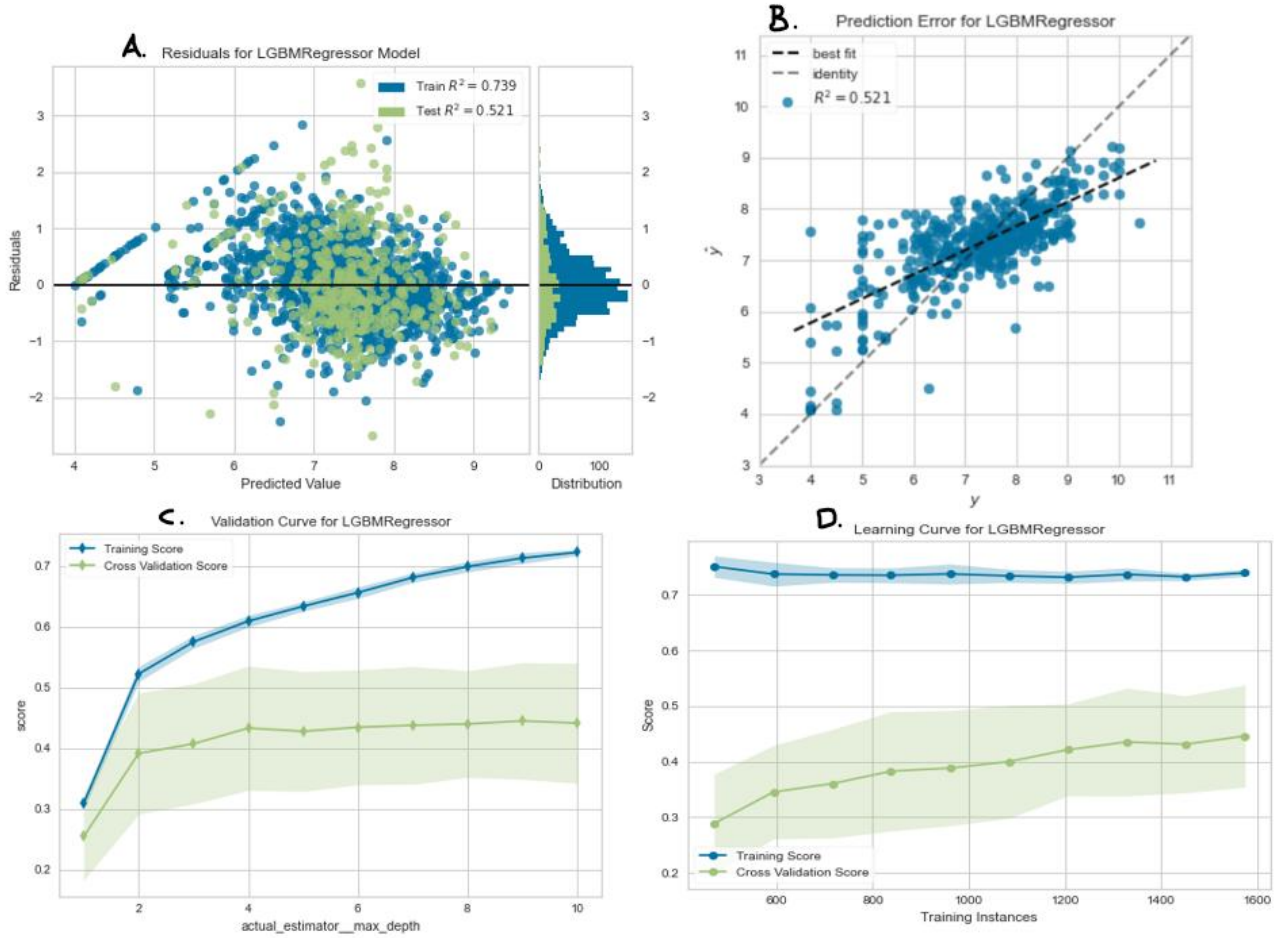


Fig.7 Performance Plots of Substructure Count Fingerprint

Y column will be shuffled such that the resulting X-Y pair will be a null pair that is expected to give rise to poor model performance seen Fig.8A while the original X-Y pairs will lead to a good performance if the model is truly robust. Fig.8A & 8B

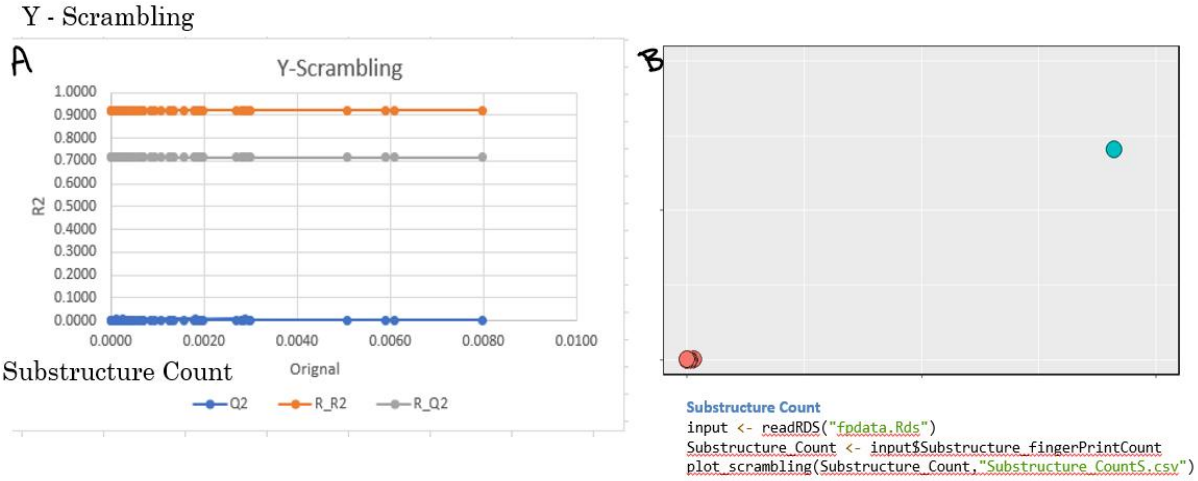
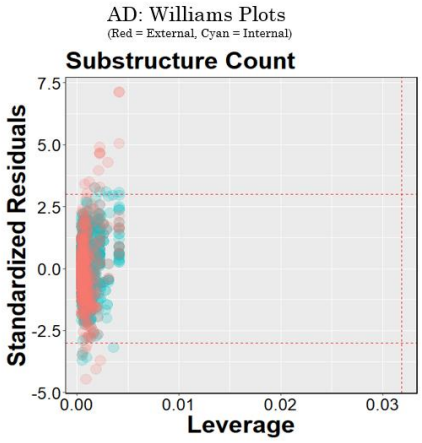


Fig. 8 Y-Scrambling Plot of Substructure Count Fingerprint

Applicability domain

The AD of the proposed QSAR model was defined as provided by the Williams plot shown in Fig. 9. The employed data set consisting of 2,186 compounds was randomly split to two separate subset in which the first subset constituting 80% of the data set was used as an internal set while the second subset constituting the remaining 20% were used as an external set. Compounds representing the internal set (blue dots) and external set (red dots) are shown in the Williams plot and it can be

clearly seen that almost all of the 2,186 compounds were located within the boundaries of applicability domain, which indicated that our proposed QSAR model had a well-defined AD. Seeing the plot at -3 consist of 5 internal compounds (66,269,761,1197,1225) and 6 External compounds (1801,1805,1873,1901,1928,1987). To see at +3 consist of 4 internal compounds (209,858,978,1443) and 14 external (1846, 1923, 1964, 968, 1998, 2056, 2115, 2120, 2133, 2137, 2138, 2150, 2163, 2184) compounds. The list can be seen in Table.4.



Compounds in the internal and external sets are shown by blue and red dots, respectively. The solid and dashed lines correspond to the ±3 standardized residual and the warning leverage value ($h^* = 0.031$), respectively

Fig.9 Williams Plots

Com pNo.	M W	lo gP	H B A	H B D	Ro BD	Structur e Type	Smiles
2137	433 .51	2. 79	7	2	5	MOL	Cc1ccc(C(=O)NC2CC2)cc1-c1cc2cnc(NC3CCOCC3)nc2n(C)c1=O
1225	435 .51	3. 55	8	3	6	MOL	CCNC(=O)c1cn2ncnc(Nc3cc(C(=O)Nc4nccs4)ccc3C)c2c1C
1197	446 .46	4. 14	5	1	4	MOL	Cc1ccc(C(=O)NC2CC2)cc1-c1cc2cnnc(-c3ccc(F)cc3F)c2n(C)c1=O
2163	382 .46	4. 36	5	1	5	MOL	CSc1nc(-c2ccc(F)cc2)c(-c2ccnc(NC(=O)C3CC3)c2)n1C
1443	359 .43	4. 83	4	2	3	MOL	O=C1Nc2cc(-c3cccc4c(NC5CC5)noc34)ccc2C12CCCC2
2138	388 .44	5. 73	3	2	4	MOL	O=C(Nc1cc(Nc2ccc3c(c2)CCCCC3=O)ccc1F)c1ccccc1
761	448 .73	5. 9	3	1	3	MOL	COc1cc(-c2ccccc2Cl)c2c(c1)N(c1c(Cl)cccc1Cl)C(=O)C(O)C2
269	459 .48	4. 59	8	1	5	MOL	Fc1ccc(-c2ncn(C3CCNCC3)c2-c2ccnc(Oc3ccc4c(c3)OCO4)n2)cc1
209	374 .51	5. 56	4	2	5	MOL	CCCOC(=O)c1sc(C(C)(C)C)cc1NC(=O)Nc1ccc(C)cc1
66	420 .3	6. 04	3	1	3	MOL	O=C1NCc2nc(Sc3ccccc3F)ccc2N1c1c(Cl)cccc1Cl
1846	499 .46	3. 05	6	4	5	MOL	COc1c(NC(=O)N[C@@H]2Cc3c(F)cccc3[C@@H]2O)cc(N2CCOCC2)cc1[C@@H](O)C(F)(F)F
1873	312 .34	1. 64	7	3	4	MOL	CCNC(=O)Nc1ccc2ncc(Nc3ccn(C)n3)nc2n1
1805	322 .37	3. 22	5	3	4	MOL	CCNC(=O)Nc1ccc2ncc(Nc3cccc(C)c3)nc2n1
1964	724 .83	7. 71	11	4	12	MOL	CC(C)c1cc(NC(=O)Nc2ccc(Oc3ccnc(Nc4ccc5[nH]ncc5c4)n3)c3ccccc23)n(-c2ccc(OCCN3CCOCC3)cc2)n1
1987	853	7. 08	15	3	14	MOL	CC(C)(C)c1cc(NC(=O)Nc2ccc(Oc3ccnc(Nc4ccc5c(cnn5CCN5CCOCC5)c4)n3)c3ccccc23)n(-c2ccc(OCCN3CCOCC3)nc2)n1

1998	575.6	5.1	7	2	10	MOL	CC(C)C[C@H](NCCc1cc(F)c(-n2c(N)c(C(=O)c3ccc(F)cc3F)ccc2=O)c(F)c1)C(=O)OC(C)(C)C
1928	627.71	8.33	9	3	7	MOL	Cc1ccc(-n2nc(C(C)(C)C)cc2NC(=O)Nc2ccc(Oc3ccnc(Nc4ccc5c(c4)OCO5)n3)c3ccccc23)cc1
1801	502.62	3.33	6	4	10	MOL	NC(=O)[C@H](CCC1CCCCC1)NC(=O)c1ccc(CNC(=O)c2cnn(-c3ccccc3)c2N)cc1
1923	658.17	9.13	8	4	7	MOL	Cc1ccc(-n2nc(C(C)(C)C)cc2NC(=O)Nc2ccc(Oc3ccnc(Nc4cc(Cl)c5[nH]ncc5c4)n3)c3ccccc23)cc1
1968	752.88	8.19	11	4	11	MOL	Cc1cc(Nc2nccc(Oc3ccc(NC(=O)Nc4cc(C(C)(C)C)nn4-c4ccc(OCCN5CCOCC5)cc4)c4ccccc34)n2)cc2cn[nH]c12
1901	624.71	7.87	9	4	7	MOL	Cc1ccc(-n2nc(C(C)(C)C)cc2NC(=O)Nc2ccc(Oc3ccnc(Nc4cnc5[nH]ncc5c4)n3)c3ccccc23)cc1
2056	503.53	5.45	5	3	5	MOL	NC(=O)c1ccc2c(c1)C(=O)c1ccc(Nc3cc(NC(=O)c4cccs4)c(F)cc3F)cc1CC2
2120	437.48	5.47	5	1	4	MOL	Cc1ccc(C(=O)Nc2cc(-c3c(-c4ccc(F)c(C)c4)nc4cccnn34)ccn2)cc1
2115	423.45	5.16	5	1	4	MOL	Cc1cc(-c2nc3cccnn3c2-c2ccnc(NC(=O)c3ccccc3)c2)ccc1F
2184	459.59	4.47	5	3	9	MOL	Nc1c(C(=O)NCc2ccc(C(=O)NCCCC3CCCCC3)cc2)cnn1-c1ccccc1
2150	436.43	5.28	4	2	5	MOL	CSc1nc(-c2ccc(F)cc2)c(-c2ccnc(NC(=O)C3(C(F)(F)F)CC3)c2)[nH]1
2133	358.41	3.23	6	2	4	MOL	CNC(=O)c1ccc(C)c(Nc2nncc3c2cnn3-c2ccccc2)c1
858	415.42	4.83	4	1	5	MOL	Cc1ccc(C(=O)NC2CC2)cc1-c1ccc2c(O[C@@H](C)C(F)(F)F)nncc2c1
978	469.88	4.29	6	2	5	MOL	Cn1c(=O)cc(Nc2cc(C(=O)NC3CC3)ccc2Cl)c2cnn(-c3ccc(F)cc3F)c21

Table. 4 Excluded Compounds from Williams plot

Mechanistic Interpretation Of Feature Importance

Feature importance analysis help reveal features that are important toward bioactivity.

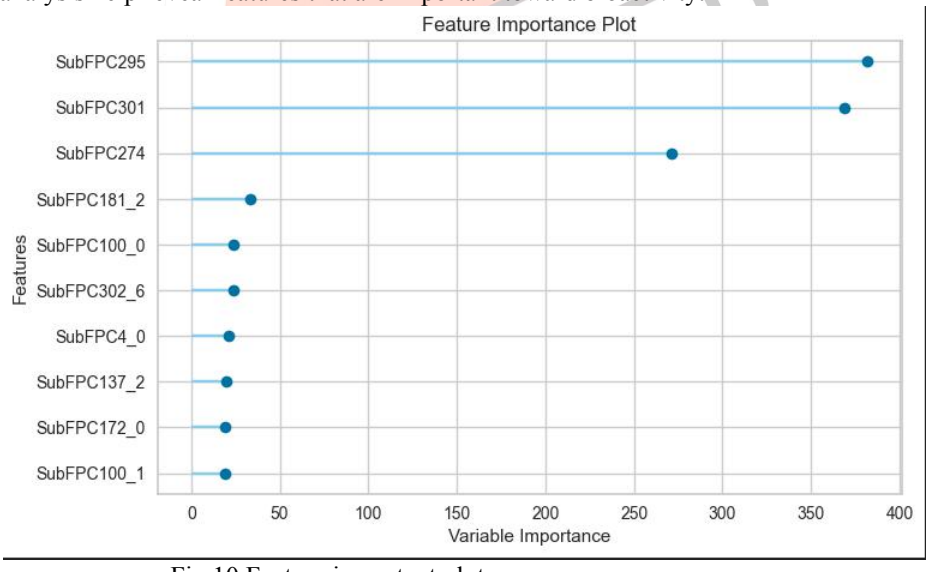


Fig.10 Feature important plot

The Fig.10 Show the important fingerprint identified by the model which give variable importance to the model. The Fig.11 is Recursive Feature Plot the maximum number of features required to capture the highest score built on 10-fold CV.

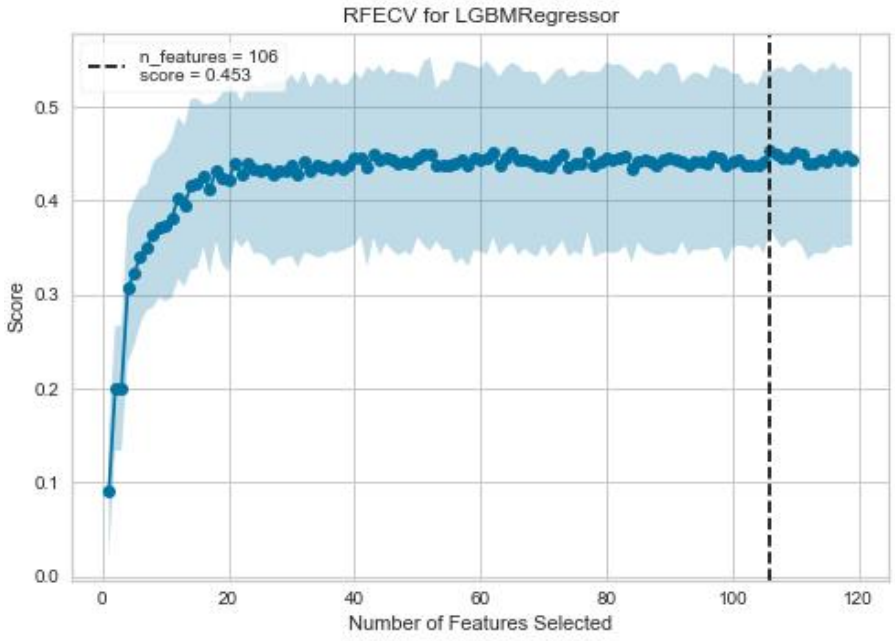


Fig.11 Recursive Feature Plot

The SHAP is library in python. SHAP summary plots adds interpretability to constructed models so that the contribution of each features to the prediction can be elucidated.Fig.12. Show the important feature and also show whether its toward active compound or inactive.

The Fig.13.Force plot and this plot essentially describes the push and pull effect that each individual features has on the base value that eventually leads to the predicted output value.

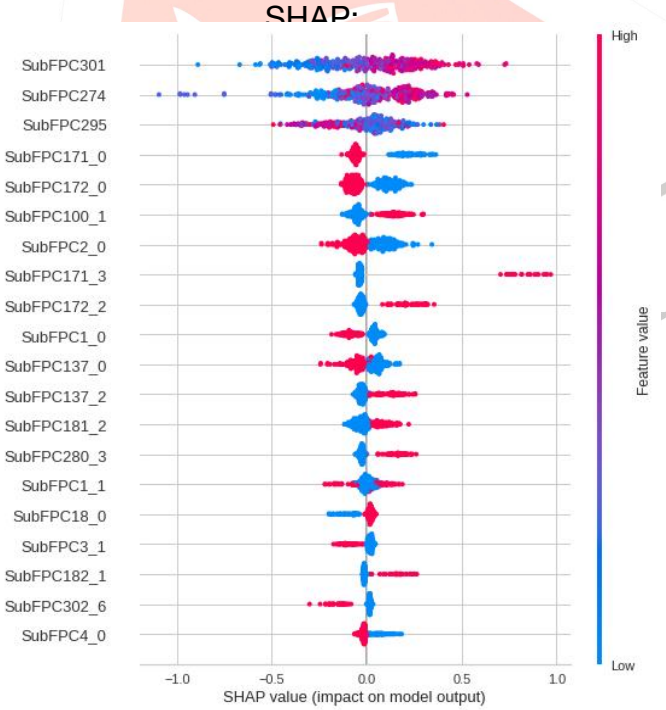


Fig.12 Feature Summary plot

Fig.13 Force Plot Showing Pulling Effect

Top Features

SubFPC301	1,5-Tautomerizable
SubFPC295	C ONS bond
SubFPC274	Aromatic ring

Other Features

SubFPC100	Secondary amide
SubFPC137	Vinylogous ester
SubFPC171	Arylchloride
SubFPC181	Hetero N nonbasic
SubFPC302	Rotatable bond

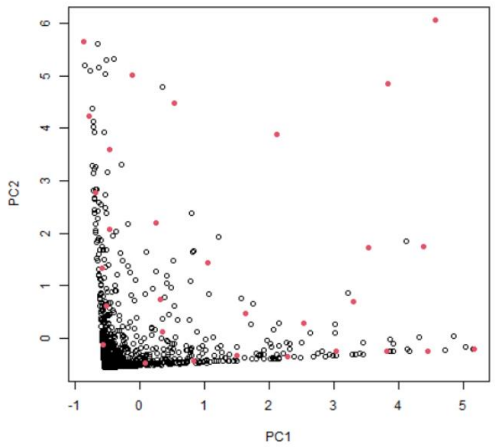
Fig.14 Top Listed Features

As Seen in Fig.10 & Fig.12 The top feature as consolidate in Fig.14

The C ONS bond (SubFPC295), which is defined as the presence of any carbon connected with either oxygen, nitrogen or sulfur atom in a molecule. These atoms are considered as high electron density atoms, which exerted from higher electronegativity comparing with a carbon atom. Unequal sharing of electron pair making covalent bond contribute polarity and afford dipole moment to a molecule, which able to generate a dipole–dipole attraction such as hydrogen bond between two polar molecules. Furthermore, increasing of polarity and presenting of hydrogen bond improve water solubility, which is essential characteristic of a drug.

The aromatic ring (SubFPC274). Findings from X-ray crystallographic study showed that in the binding site of the co-crystal structure of MAPK aromatic ring engages in a π–π stacking interaction with the PHE 169 (human p38 MAPKα numbering), thereby indicating the importance of the aromatic ring for p38 MAPKα inhibition.

SubFPC301 (1,5-Tautomerizable). Tautomerizable heterocycles have recently emerged as an attractive class of inhibitor.



Distribution of active p38 MAPKα inhibitors (gray circles) and the diversity set (red circles) selected for molecular docking.

Fig.15 Kennard-stone plot

Title	ID	Smiles	BioActive	IC50 nM	kacl/mol
ligprep_4.sdf:23	CHEMBL429582	Cc1ccc(NC(=O)c2cccc(	active	1	-13.021
ligprep_4.sdf:3	CHEMBL215969	CNc1ncc2cc(-c3cc(C(=	active	50	-12.859
ligprep_4.sdf:16	CHEMBL39931	Cn1nc(C(C)(C)C)cc1N	active	290	-10.521
ligprep_4.sdf:22	CHEMBL272535	Fc1cccc(F)c1-c1nc(-c2	active	88.4	-10.239
ligprep_4.sdf:5	CHEMBL289155	CC(=O)Nc1ccc(Oc2ccc	active	235	-10.015
ligprep_4.sdf:15	CHEMBL477583	Cc1ccc(NC(=O)c2ccoc	active	16	-9.979
ligprep_4.sdf:10	CHEMBL19086	Cc1ccc(NC(=O)Nc2cc(	active	741	-9.907
ligprep_4.sdf:26	CHEMBL278275	CCCOC(=O)c1sc(C(C)(	active	482	-9.907
ligprep_4.sdf:7	CHEMBL478650	Cc1ccc(C(=O)Nc2nncs	active	150	-9.831
ligprep_4.sdf:24	CHEMBL215943	Cc1ccc(C(=O)Nc2cccc(	active	3	-9.789
ligprep_4.sdf:13	CHEMBL18709	COC(=O)c1c(NC(=O)N	active	663	-9.55
ligprep_4.sdf:28	CHEMBL215680	CCc1nnc2ccc(-c3ocnc	active	54.5	-9.407
ligprep_4.sdf:8	CHEMBL202666	CC(=O)N1CCC[C@H](	active	162	-9.252
ligprep_4.sdf:18	CHEMBL62665	CC(C)(C)c1cc(NC(=O)I	active	710	-9.146
ligprep_4.sdf:25	CHEMBL262866	CCNC(=O)c1cn2ncnc(I	active	16	-8.776
ligprep_4.sdf:17	CHEMBL155687	C[S+]([O-])c1ccc(-c2c	active	670	-8.635
ligprep_4.sdf:30	CHEMBL279560	COC(=O)c1sc(C(C)(C)(	active	610	-8.205
ligprep_4.sdf:21	CHEMBL402943	CCNC(=O)c1ccc(C)c(N	active	330	-8.184
ligprep_4.sdf:19	CHEMBL372384	CCc1nc(-c2cccc(C)c2)c	active	1000	-8.09
ligprep_4.sdf:11	CHEMBL552375	Nc1c(C(=O)c2ccc(F)cc	active	24	-7.809
ligprep_4.sdf:27	CHEMBL516908	Cc1ccc(NC(=O)CC2CC	active	450	-7.361
ligprep_4.sdf:6	CHEMBL551189	COc1cc(F)c(-c2c(N)c((	active	5	-7.233
ligprep_4.sdf:14	CHEMBL261005	CNC(=O)c1cn2ncnc(N	active	4	-6.968
ligprep_4.sdf:12	CHEMBL437024	NCC(=O)NCCOc1cc(-c	active	80	-6.885
ligprep_4.sdf:1	CHEMBL559597	CCn1c(=O)cc(Nc2cc(C	active	4	-6.783
ligprep_4.sdf:4	CHEMBL94783	COC(=O)c1cc(-c2ccc(F	active	11	-6.663
ligprep_4.sdf:20	CHEMBL94556	COC(=O)c1cc(-c2ccccc	active	1	-6.277
ligprep_4.sdf:2	CHEMBL243980	CN(C)C(=O)C(=O)c1cr	active	195	-5.636
ligprep_4.sdf:9	CHEMBL262608	CCNC(=O)c1cn2ncnc(I	active	2	-4.874
Exclusion SET					

Fig.16 Exclusion Set

Molecular Docking of p38 MAPKα Inhibitors

To gain a further understanding on the non-covalent interaction between p38 MAPKα and their inhibitors, a chemically diverse Exclusion set of 30 representative compounds was extracted from active p38 MAPKα inhibitors (i.e., having IC_{50} <1 μM) using the Kennard–Stone algorithm and subjected to an investigation on its binding modality against the active site of p38 MAPK. Fig.15 shows the distribution of the selected subset of compounds in the context of the full set of actives, which was found to provide a full coverage of the original chemical space. The formation of this binding site requires a large conformational change not observed previously for any of the protein Ser/Thr kinases. This change is in the highly conserved Asp-Phe-Gly motif within the active site of the kinase. Solution studies demonstrate that this class of compounds has slow binding kinetics, consistent with the requirement for conformational change. Improving interactions in this allosteric pocket, as well as establishing binding interactions in the ATP pocket (Pargellis et al., 2002) Consequently, the binding modality was analyzed in order to gain understanding on the contribution of key residues in interacting with the investigated set of 30 compounds. This was performed using the SiMMap web server, which revealed three major binding anchors: vdW1 and vdW2 along with their site-moiety preferences. ligand moieties: aromatic ring, Heterocyclic moiety, alkenes Furthermore, analysis of the binding energy from the 30 representative compounds revealed that compounds **23**, **3** and **16** exhibited the lowest binding energy of -13.020, -12.859 and -10.521 respectively

Ligand Interaction Plot Compound 23 Fig. 17

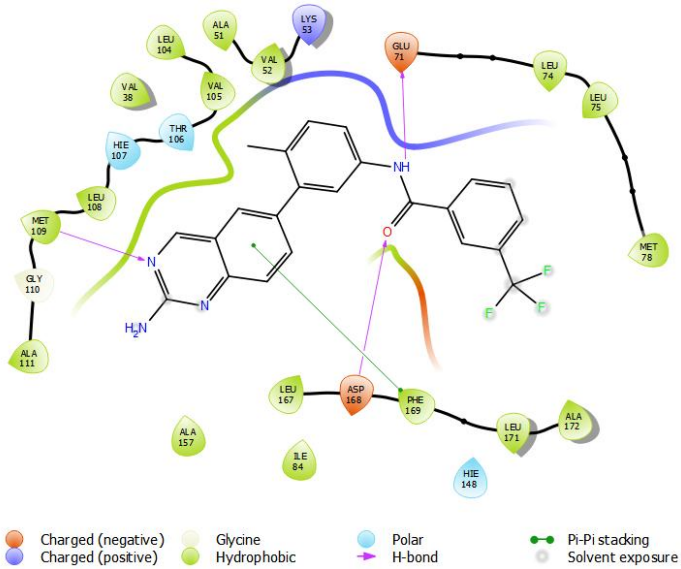


Fig.17 Compound 23

Ligand Interaction Plot Compound 3 Fig. 18

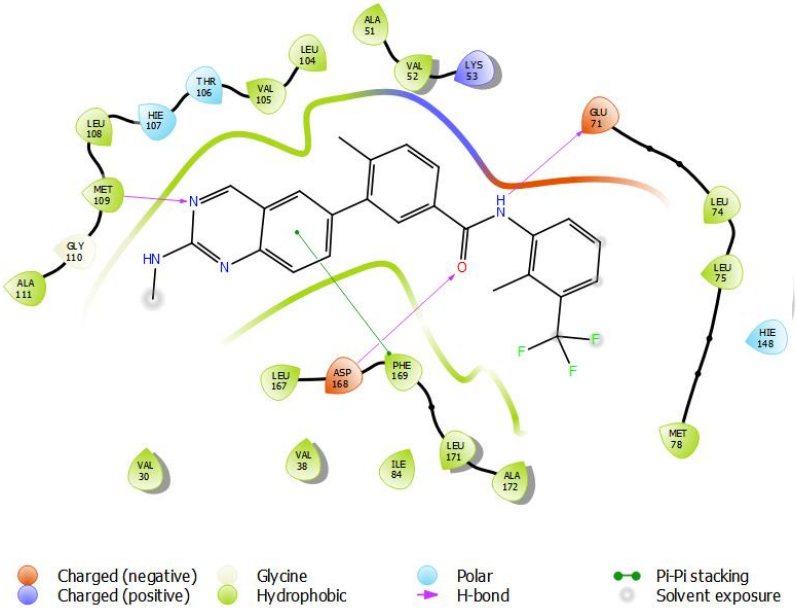


Fig.18 Compound 3

Ligand Interaction Plot Compound 16 Fig.19

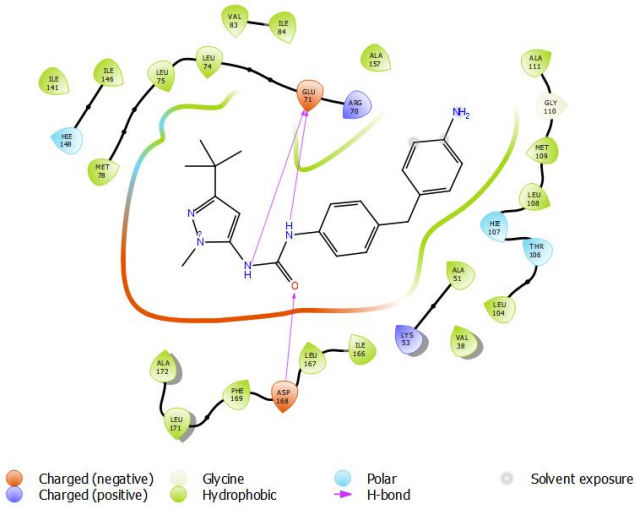


Fig.19 Compound 16

Compound 23 Fig.17 show a hydrogen bond interaction with GLU71, ASP168, MET109. Wherein the PHE169 has pi-pi interaction at the aromatic site.
Compound 3 Fig.18 same interaction as Compound 23 but with different conformer.
Compound 16 Fig.19 duel hydrogen bond interaction with GLU71 and an another with ASP168 Interestingly, this finding is corroborated by the feature importance results obtained from the QSAR model as shown in Fig.14 in which the aromatic moiety, C ONS bond, secondary, heterocyclic ring and the hetero N non-basic moiety were found amongst the important substructures that are essential for the bioactivity of p38 MAPKα inhibitor.

CONCLUSION

Twelve sets of fingerprint descriptors were used for constructing QSAR models and their performances were comparatively evaluated. It was observed that several fingerprint descriptors afforded good performance for the constructed models indicating that they could capture the feature space of p38 MAPs inhibitors. By taking advantage of the built-in feature importance estimator and SHAP Library plots, the following important features that are for MAPKs inhibition were identified: **SubFPC301(1,5-Tautomerizable), SubFPC295(C ONS bond), SubFPC274(Aromatic ring) and SubFPC181(Hetero N nonbasic).** Results from molecular docking also support the aforementioned findings from the QSAR models in which the aromatic, and heteroaromatic rings were preferable moieties for interacting with the hydrophobic pocket of p38 MAPKα . It is anticipated that the knowledge gained from this study could be used as general guidelines for the design of novel p38 MAPKα inhibitors.

REFERENCES

1. Global Initiative for Asthma. Glo12bal Strategy for Asthma Management and Prevention, 2021. Available from: [ginasthma.org]
2. A Valero; S Quirce; I Dávila; J Delgado; J Domínguez-Ortega. Allergy. 2017
3. AL Durham; G Caramori; KF Chung; IM Adcock. Translational Res. 2016, 167(1), 192-203.
4. Karpakavalli M, Sangilimuthu AY , Sasikala M, Komala M , Ranjithkumar D and Mohan S J. Chem. Pharm. Res., 2017, 9(1):43-44
5. Kian Fan Chung,p38 Mitogen-Activated Protein Kinase Pathways in Asthma and COPD,Chest,Volume 139, Issue 6,2011,Pages 1470-1479,ISSN 0012-3692,https://doi.org/10.1378/chest.10-1914
6. Jiang, Y., Gram, H., Zhao, M., New, L., Gu, J., Feng, L., Di Padova, F., Ulevitch, R. J., & Han, J. (1997). Characterization of the structure and function of the fourth member of p38 group mitogen-activated protein kinases, p38delta. J Biol Chem, 272, 30122-30128
7. Federico Mayor, Jr., Maria Jurado-Pueyo, Pedro M. Campos & Cristina Murga (2007) Interfering with MAP Kinase Docking Interactions: Implications and Perspectives for the p38 Route, Cell Cycle, 6:5, 531, DOI: 10.4161/cc.6.5.3920
8. Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F., Bellis, L., & Overington, J. P. (2015). ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research*, *43*(W1), W612–W620. doi.org/10.1093/nar/gkv352
9. Suvannang, Naravut and Preeyanon, Likit and Malik, Aijaz Ahmad and Schaduangrat, Nalini and Shoombuatong, Watshara and Worachartcheewan, Apilak and Tantimongcolwat, Tanawut and Nantasenamat, Chanin (2018) "Probing the origin of estrogen receptor alpha inhibition via large-scale QSAR study", RSC Adv, 8:21, 11322-11356, DOI: doi.org/10.1039/C7RA10979B
10. Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. 2009. A practical overview of quantitative structure–activity relationship. EXCLI Journal 8(7):74-88
11. Chu EK, Drazen JM. Asthma: one hundred years of treatment and onward. Am J Respir Crit Care Med. 2005;171(11):1202–8. https://doi.org/10.1164/rccm.200502-257OE.
12. Crompton, G. A brief history of inhaled asthma therapy over the last fifty years. *Prim Care Respir J* **15,** 326–331 (2006). https://doi.org/10.1016/j.pcrj.2006.09.002
13. Pelaia C, Vatrella A, Gallelli L, et al. Role of p38 Mitogen-Activated Protein Kinase in Asthma and COPD: Pathogenic Aspects and Potential Targeted Therapies. Drug Design, Development and Therapy. 2021 ;15:1275-1284. DOI: 10.2147/dddt.s300988.
14. Yang, Y., S. C. Kim, T. Yu, Y.-S. Yi, M. H. Rhee, et al. 2014. Functional roles of p38 mitogen-activated protein kinase in macrophage-mediated inflammatory responses. Mediators Inflammatory Responses 2014: 352371–352371 DOI: doi.org/10.1155/2014/352371
15. Mohan Babu Jatavath, Sree Kanth Sivan, Yamini Lingala And Vijjulatha Manga, Docking and 3D QSAR Studies on p38α MAP Kinase Inhibitors,2011, 8(4), 1596-1605 core.ac.uk/download/pdf/192698212.pdf
16. Bachstetter, Adam & Van Eldik, Linda. (2010). The p38 MAP Kinase Family as Regulators of Proinflammatory Cytokine Production in Degenerative Diseases of the CNS. Aging and disease. 1. 199-211.
17. Pat Bass, MD, What Type of Asthma Do You Have? (verywellhealth.com), 2021
18. WHO WHO | Causes of asthma, 2020
19. Pichardo, Asthma: Causes, Symptoms, Diagnosis, Treatment (webmd.com), 2019
20. David S Millan, Future Medicinal Chemistry 2011 3:13, 1635-1645
21. P. Bhavsar, N. Khorasani, M. Hew, M. Johnson, K. F. Chung European Respiratory Journal Apr 2010, 35 (4) 750-756; DOI: 10.1183/09031936.00071309

22. Dr. Peter Schochet and Dr. Hauw Lie Severe Asthma in Children – Pediatric Pulmonologists in Plano (pedilung.com)
23. MAYO, Asthma - Diagnosis and treatment - Mayo Clinic, 2020
24. Safavi, Reihaneh & Ghasemi, Jahan. (2013). Application of 3D-QSAR on a Series of Potent P38-MAP Kinase Inhibitors. Journal of Applied Chemical Research. 7. 64-74.
25. Bhansali S, Kulkarni VM. Pharmacophore generation, atom-based 3D-QSAR, docking, and virtual screening studies of p38-α mitogen activated protein kinase inhibitors: pyridopyridazin-6-ones (part 2). *Research and Reports in Medicinal Chemistry*. 2014;4:1-21
https://doi.org/10.2147/RRMC.S50738
26. El Ghalia Hadaji, Mohamed Bourass, Abdelkarim Ouammou, Mohammed Bouachrine,3D-QSAR models to predict anti-cancer activity on a series of protein P38 MAP kinase inhibitors,Journal of Taibah University for Science,Volume 11, Issue 3,2017,Pages 392-407,ISSN 1658-3655,https://doi.org/10.1016/j.jtusci.2016.05.006.
27. pycaret.org. PyCaret, April 2020. URL https://pycaret.org/about. PyCaret version 2.3.1(Resource)
28. Yap, C.W. (2011), PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem., 32: 1466-1474. https://doi.org/10.1002/jcc.21707
29. Gaulton et al *Nucleic Acids Research*, Volume 40, Issue D1, 1 January 2012, Pages D1100–D1107, https://doi.org/10.1093/nar/gkr777
30. R. W. Kennard & L. A. Stone (1969) Computer Aided Design of Experiments, Technometrics, 11:1, 137-148, DOI: 10.1080/00401706.1969.10490666
31. Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen Journal of Chemical Information and Computer Sciences 2003 43 (2), 493-500,DOI: 10.1021/ci025584y
32. Lowell H. Hall and Lemont B. Kier,Journal of Chemical Information and Computer Sciences 1995 35 (6), 1039-1045,DOI: 10.1021/ci00028a014
33. Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse,Journal of Chemical Information and Computer Sciences 2002 42 (6), 1273-1280,DOI: 10.1021/ci010132r
34. NCBI. 2009. PubChem Substructure Fingerprint. Version 1.3
35. Justin Klekota, Frederick P. Roth, Chemical substructures that enrich for biological activity, *Bioinformatics*, Volume 24, Issue 21, 1 November 2008, Pages 2518–2525, https://doi.org/10.1093/bioinformatics/btn479
36. Raymond E. Carhart, Dennis H. Smith, and R. Atom pair Venkataraghavan,Journal of Chemical Information and Computer Sciences 1985 25 (2), 64-73.DOI: 10.1021/ci00046a002
37. Kuhn M. 2008.Building predictive models in R using the caret package. Journal of Statistical Software 28(5):1-26
38. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.
39. Simeon S, Anuwongcharoen N, Shoombuatong W, Malik AA, Prachayasittikul V, Wikberg JES, Nantasenamat C. 2016. Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ* 4:e2322 https://doi.org/10.7717/peerj.2322
40. Schrödinger Release 2018-4: Maestro, Schrödinger, LLC, New York, NY, 2021.
41. Bollback, J.P. SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics* **7,** 88 (2006). https://doi.org/10.1186/1471-2105-7-88
42. Github royalseeker/p38_MAPK_ASTHMA_QSAR: Project Files (github.com)
43. Guolin Ke, et al. in the 2017 paper titled "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." LINk
44. Pargellis, C., Tong, L., Churchill, L., Cirillo, P.F., Gilmore, T., Graham, A.G., Grob, P.M., Hickey, E.R., Moss, N., Pav, S., Regan, J.(2002) Nat Struct Biol 9: 268-272, DOI: 10.1038/nsb770