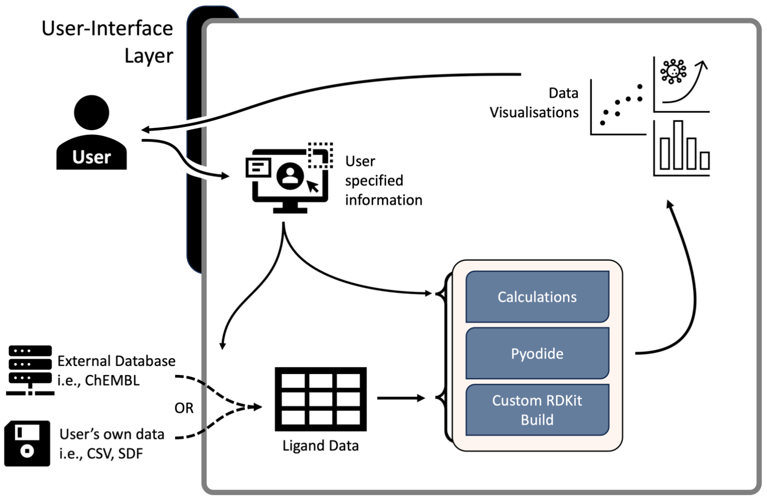
# QITB: An interactive open-source static web app for cheminformatics

Syed Zayyan Masud[1], Theo Redfern-Nichols[1], and Graham Ladds[1]
[1]Department of Pharmacology, University of Cambridge, UK

## Abstract

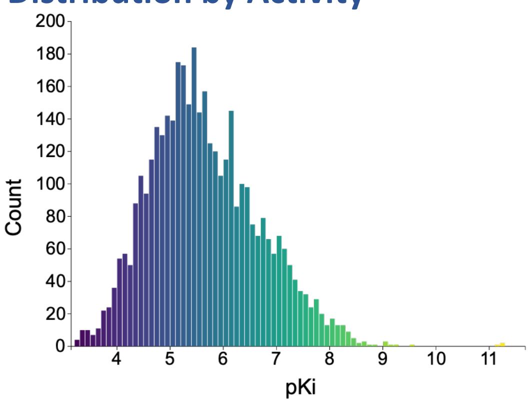
Cheminformatics has been studied in various forms for almost half a century. During these decades, innovative methods have accelerated drug discovery, while simultaneously reducing the cost. However, programming know-how remains a barrier to entry for most researchers. Therefore, we introduce a static web app for Quantitative Structure-Activity Relationship (QSAR)-in-the-browser, QITB. This app runs all cheminformatics functions solely on the consumer's device, with no external server attached and, on any device, capable of running any modern web browser. QITB allows users to fetch data from external services like ChEMBL or load their data. Following this, data pre-processing, interactive chemical space visualisation and QSAR models allow the user to analyse their molecules to easily search for patterns and predict the activity of novel compounds. The code for this data is Open Source and the web app itself is hosted through GitHub Pages.



## Molecular Data Visualisation

Researchers may visualise their data or import data for a particular target using the ChEMBL database. Following target selection, data is automatically processed and can be visualised in a variety of ways. Ligands with binding affinity for Adenosine A1 Receptor, obtained from ChEMBL, shall be used to showcase these various visualisations.

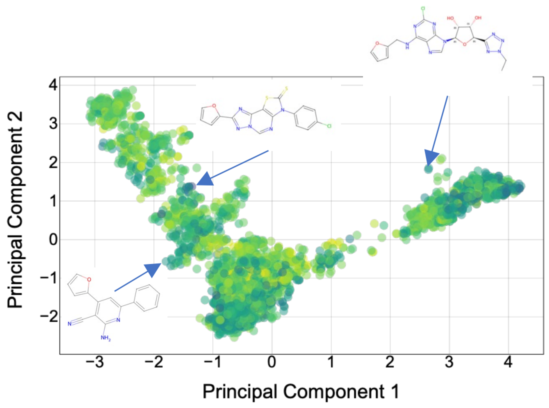
### Distribution by Activity



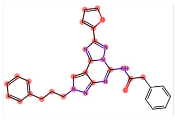
Data can be visualised using an activity distribution histogram. When the activity is binding, the x-axis shows (pKi) of small molecules to A1R, and the y-axis depicts the number of small molecules in each pKi range. This is useful to detect if the data is normally distributed and observe abnormalities.

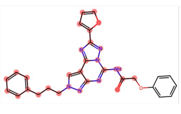
### Dimensionality Reduction

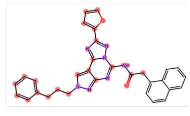
Conversion of SMILES into Morgan fingerprints (ECFP4), allows principal component analysis (PCA) to organise molecules based on their similarity to each other. Here, axes indicate the two reduced dimensions from the PCA calculation. This could be useful to observe the diversity in the data and to recognise possible trends with the activity measurement of interest.

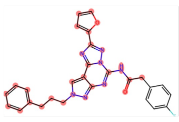


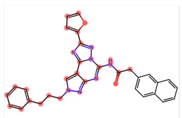
### Matched Molecular Series

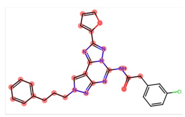

pKi : 6.12


pKi : 6.3


pKi : 4.96


pKi : 5.14


pKi : 5.62


pKi : 5.22

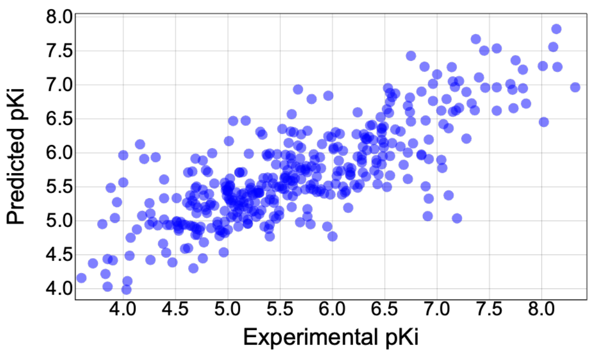
The Matched Molecular Series technique allows identification of molecules differing by only single chemical transformations. This can be useful to identify activity cliffs, where small molecular changes can lead to large changes in activity (pKi in this case). [1]

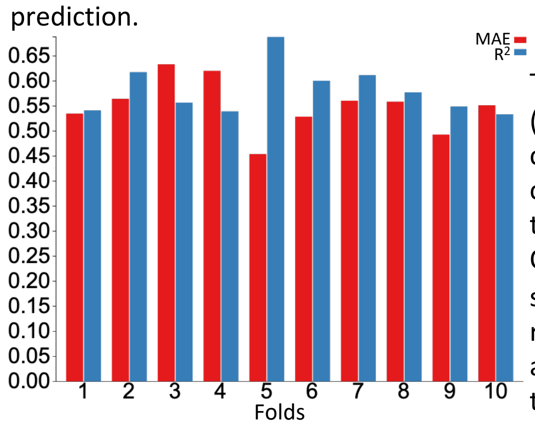
## Machine Learning

QITB allows for machine learning-based prediction of ligand activity. Using the Scikit-Learn library, a Random Forest model is trained on the web browser using Pyodide. The model is evaluated with a 10-fold, using the A1R ChEMBL data as input with Morgan (ECFP4) Fingerprints.

### Model Performance Evaluation

Once all the folds are tested, each can be investigated with a correlation plot. The x-axis shows the experimental activity (pKi) from the test set data, while the y-axis represents the model's prediction.





The Mean Absolute Error (MAE), and the correlation coefficient $R^2$ can be calculated for each fold from the correlation plot above. QITB then displays these statistics ($R^2$ in blue, MAE in red) for all folds. This helps assess the performance of the models.

### Virtual Screening

Once trained, a model can be used to predict activity for batches of molecules with a user-provided CSV file. The screened molecules can be further sampled using a Genetic Algorithm with a function named Coverage Score. [2]

## Future Directions

- QSAR models have been developed using various deep learning approaches with remarkable results i.e., chemprop. These models could be ported to QITB, and this is possible with the help of projects like the ONNX Web Runtime and TensorFlow JS.
- Currently, QITB only includes access to ChEMBL. In the future, other databases could be added i.e. PubChem, ZINC, DrugBank, BindingDB or GPCRdb.
- More file formats like SDF and PDB files will to be supported for user-provided data.

## References

[1] Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath ACS Omega 2019 4 (11), 14360-14368 DOI: 10.1021/acsomega.9b02221
[2] Daniel J. Woodward, Anthony R. Bradley, and Willem P. van Hoorn *Journal of Chemical Information and Modeling* **2022** *62* (18), 4391-4402 DOI: 10.1021/acs.jcim.2c00258

The QR code links to the source code for the website.