Predictive global models of cruzain inhibitors with large chemical coverage

Jose Guadalupe Rosas-Jimenez,[†,¶] Marco A. Garcia-Revilla,[†] Abraham Madariaga-Mazon,[‡] and Karina Martinez-Mayorga[*,‡]

†*Division de Ciencias Naturales y Exactas, Universidad de Guanajuato, Guanajuato, Mexico*

‡*Instituto de Quimica, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico*

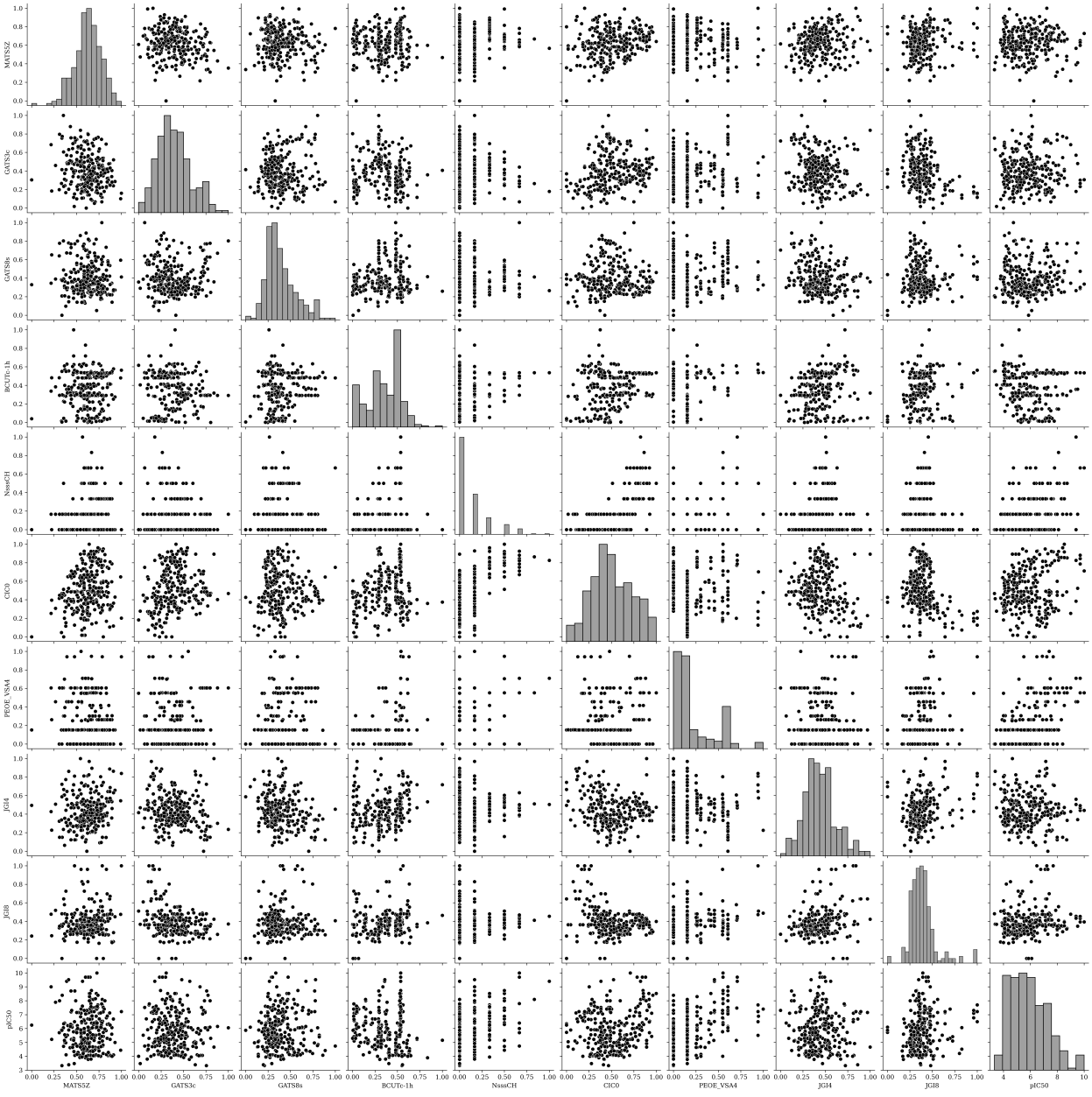
¶*Current address: Instituto de Quimica, Universidad Nacional Autonoma de Mexico, Mexico City, Mexico*

E-mail: kmtzm@unam.mx

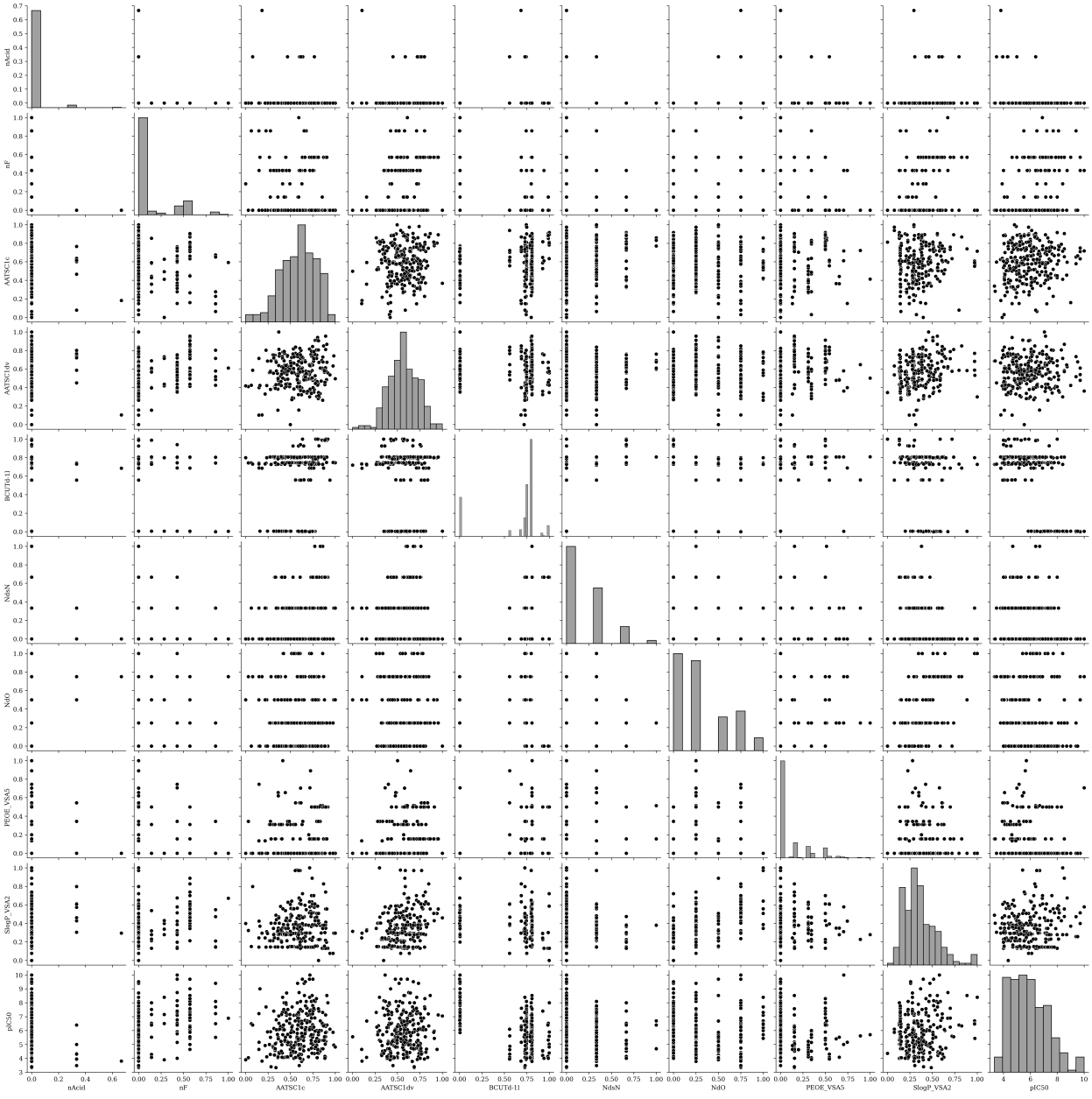
Supporting Information

Content

S1 Figure S1. Scatter plots matrix of molecular descriptors for KNN model



S2 Figure S2. Scatter plots matrix of molecular descriptors for Random Forest model



S3 Python script usage

The script is initiated from the terminal using the command:

```
python qsar_cruzain_predictor.py -i input.sdf -o output.csv
```

The argument *input.sdf* is replaced by the name of the sdf file containing the 2D molecular structures of compounds to be predicted. Further requirements and recommendations can be found in the `README` file.

Internally, the script reads the chemical structures and calculates the molecular descriptors listed in tables 6 and 7 of the main text. Then, both the KNN and RF models are applied to predict the pIC_{50} values for each compound. To analyze if the molecule falls into the applicability domain, leverage values and maximum Tanimoto similarity with the training set are also calculated. Results are saved as a comma-separated-values (csv) file and its name can be specified with the argument *output.csv* in the command line. In this file, the first 18 columns store the values for the molecular descriptors and their headers are the names listed in tables 6 and 7 of the main text. The last columns are described in table S1. The leverage and similarity values must be taken into account to consider a prediction as reliable. Columns 21, 22 and 24 can be used to generate plots similar to those in figure 11 of the main text. Only the predicted values for molecules in the region between the dashed lines should be taken with confidence.

Table S1: **Description of the output file generated by the Python script for the prediction of cruzain pIC_{50} values.**

Column	Header	Description
1–18	Descriptor names	Descriptor values for the molecules
19	pIC50_KNN	Predicted pIC_{50} using the KNN model
20	pIC50_RF	Predicted pIC_{50} using the Random Forest model
21	Max Similarity	Highest Tanimoto similarity between the query molecule and those in the training set
22	Leverage KNN	Leverage values in the KNN model descriptor space
23	Max Leverage KNN	Leverage threshold for the KNN model
24	Leverage RF	Leverage values in the RF model descriptor space
25	Max Leverage RF	Leverage threshold for the RF model
26	ID	Molecule identifiers as in the sdf file