



Understanding Toxicity Predictions Through Explainable AI

Shubham Thakur,
Supervisor : Prof. Dr. Gerhard Ecker
Mentors : Ece Asilar, Jennifer Hemmerich
Pharmacoinformatics Research Group
University of Vienna

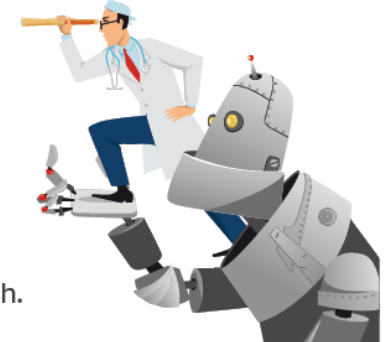
AI Success Stories

AI is helping people from everywhere to solve problems in exciting new ways



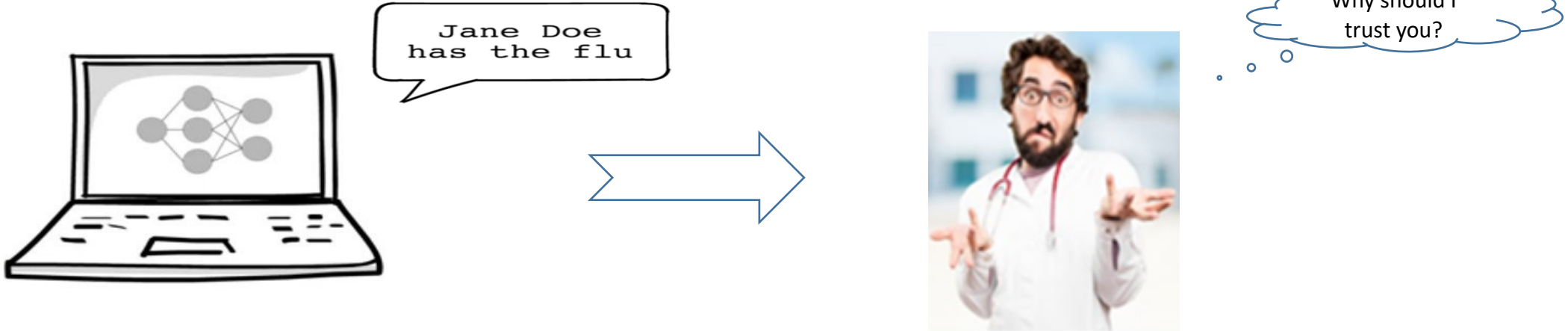
AI could save lives and improve medical diagnoses

Artificial intelligence in healthcare must combat the hype that inflates the technology's potential. But there are success stories about what AI algorithms could accomplish.

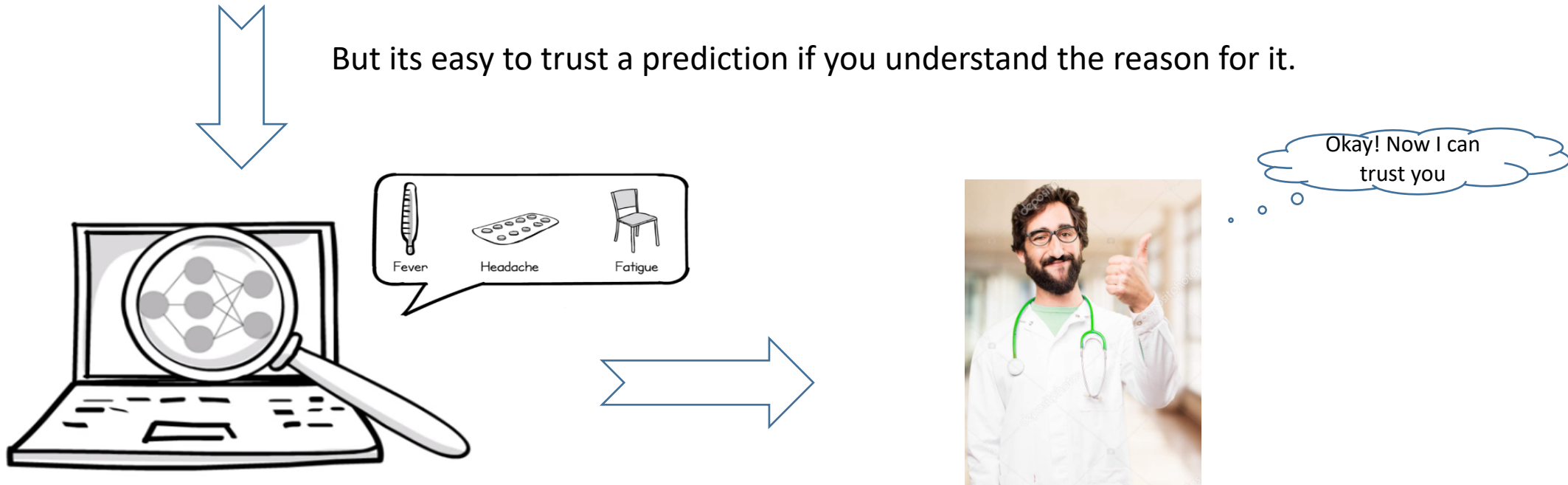


Why Explainable AI ?

Sometime you don't know if you can trust an AI model prediction.

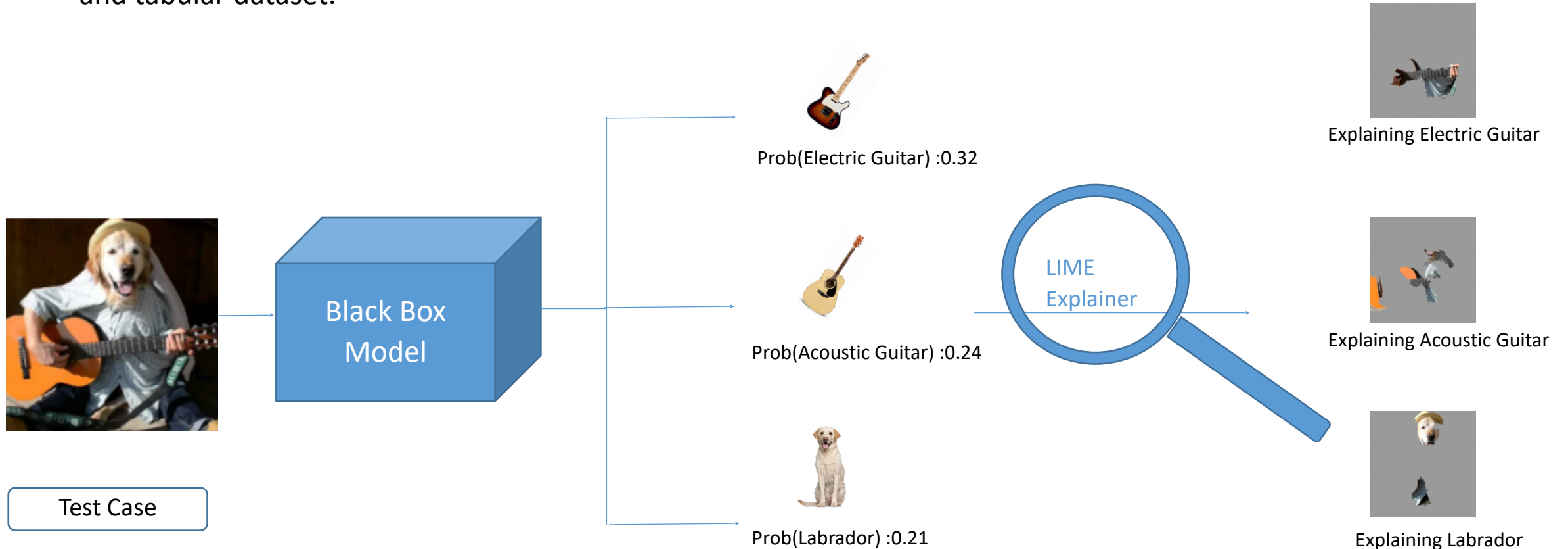


But its easy to trust a prediction if you understand the reason for it.



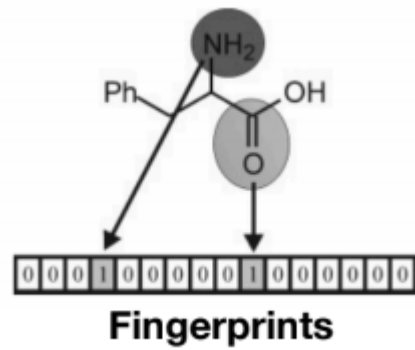
Available Tools For Explainable AI

- **SHAP**: Easy to Implement but not suitable for our data.
- **LRP(Layer-Wise Relevance Propagation)** : Robust but time consuming, we leave it for the future
- **LIME(Local Interpretable Model-Agnostics Explanation)** : Easy to Implement and can be applied to both Image and tabular dataset.

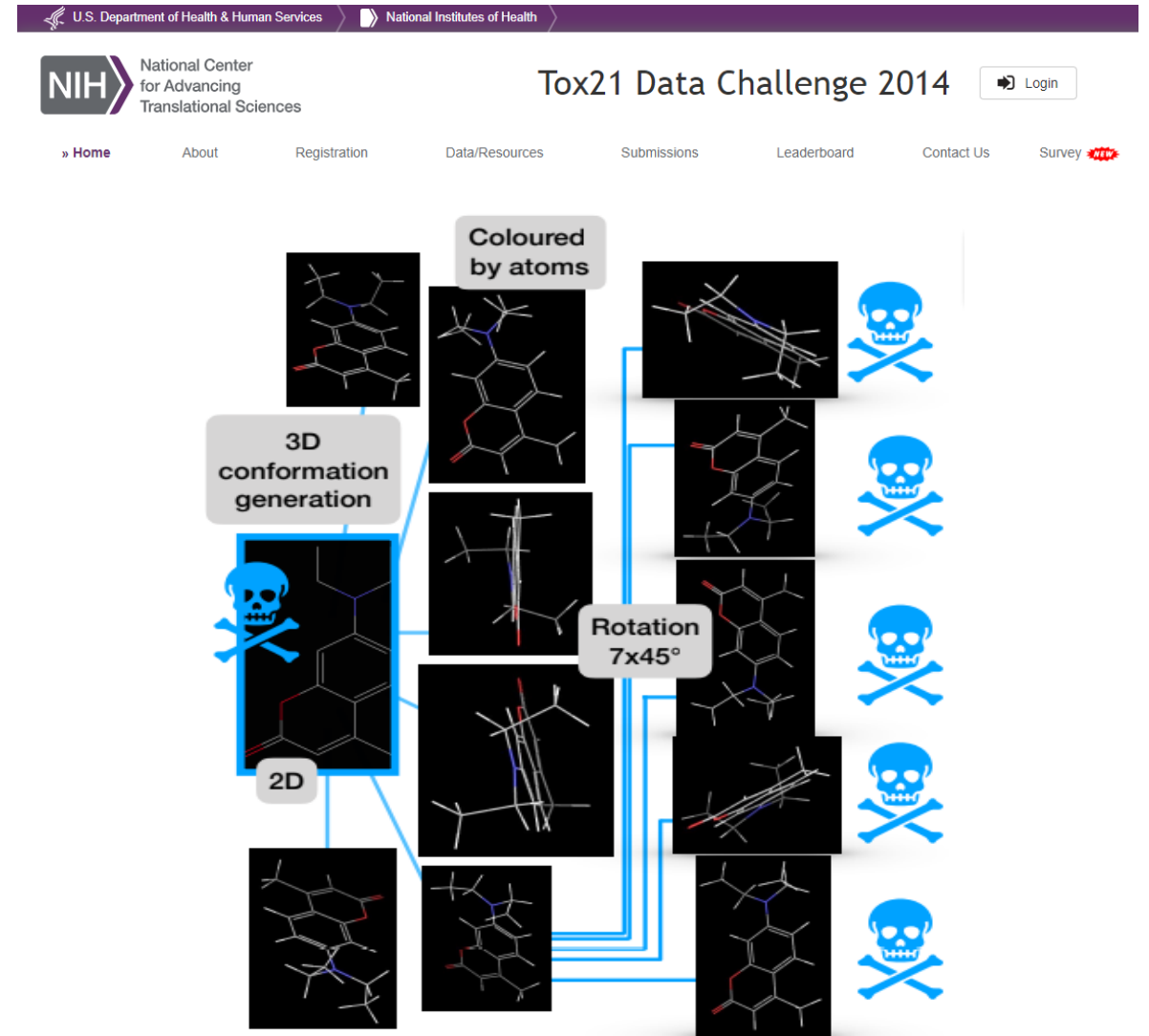


Datasets And Model

Tox-Alerts



| Reference | Activity | TA249 | TA311 | TA441 | TA470 | TA473 | TA854 | TA1095 |
|-----------|----------|-------|-------|-------|-------|-------|-------|--------|
| AAKJLRGG | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| AAOVKJBE | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| AAQOQKC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AAXVEMN | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ABJKWBD | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ACGUYXC | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| ACTIUHUL | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ACTRVOB | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ACWBQPM | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

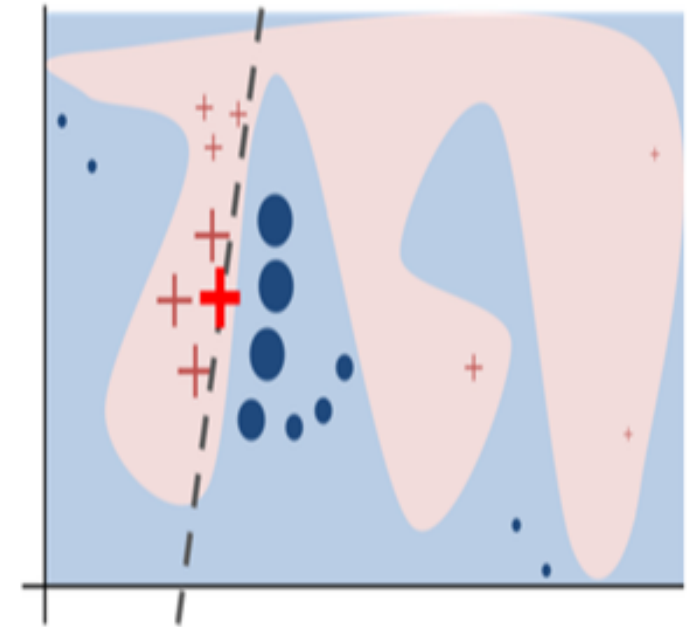


Working Principal Of LIME



With LIME we try to explain the closest decision boundary from the test case and decision involved for the same.

- Step 1 : Permute N data points.
- Step 2 : Calculate the similarity scores
- Step 3 : Make prediction on this permuted data using black box model.
- Step 4 : Fit simple linear model on the permuted data weighted by similarity scores.
- Step 5 : Extract the feature weights from coefficients of simple linear model and use this as explanation for black box models local behaviour.



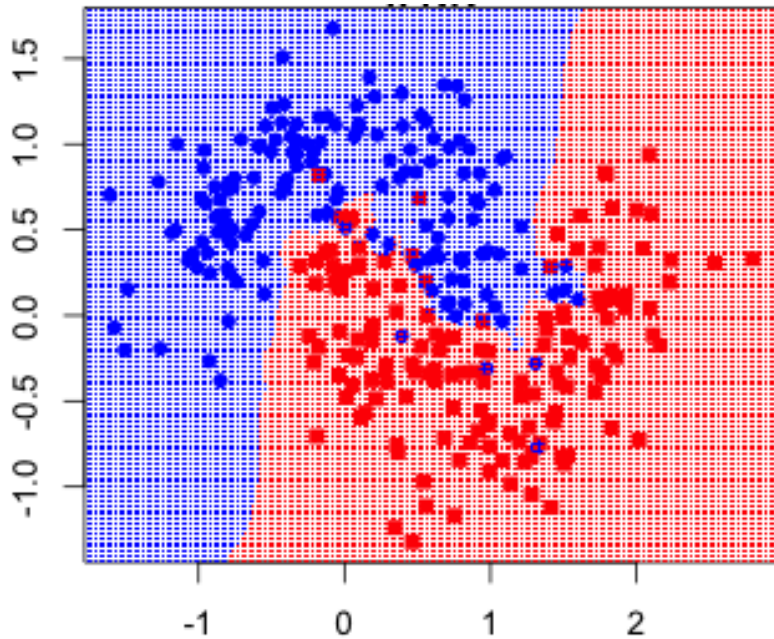
Similarity Score Calculations

| | | | | | | |
|-----|---|---|---|---|---|---|
| A : | 1 | 1 | 0 | 0 | 0 | 1 |
| B : | 0 | 1 | 0 | 1 | 0 | 1 |

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\Rightarrow \frac{\begin{array}{|c|c|c|c|c|c|} \hline 0 & 1 & 0 & 0 & 0 & 1 \\ \hline \end{array}}{\begin{array}{|c|c|c|c|c|c|} \hline 1 & 1 & 0 & 1 & 0 & 1 \\ \hline \end{array}} = \frac{2}{4}$$

Non Linear Local Decision Boundary

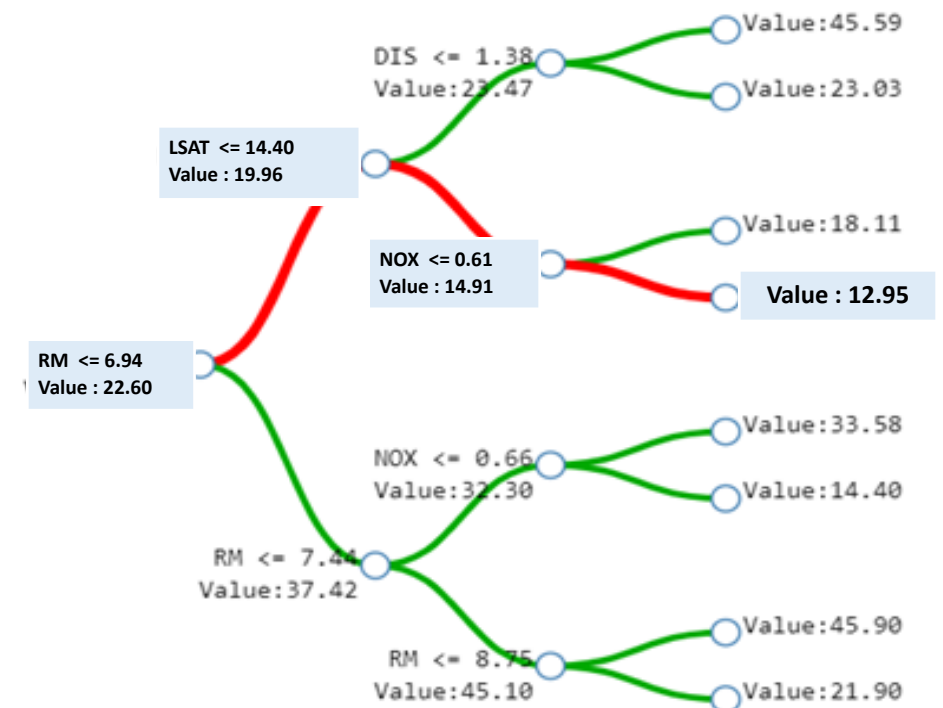


Tree interpreter

- Each prediction is decomposed into a sum of contributions from each feature.

$$\text{prediction} = \text{bias} + \text{feature}_1\text{contribution} + \dots + \text{feature}_n\text{contribution}$$

- Decision tree as the local model to capture non-linearity.
- Used tree interpreter for calculating feature importance.



Prediction: **12.95** \approx **22.60** (trainset mean) - **2.64**(loss from RM) - **5.04**(loss from LSTAT) - **1.96**(loss from NOX)

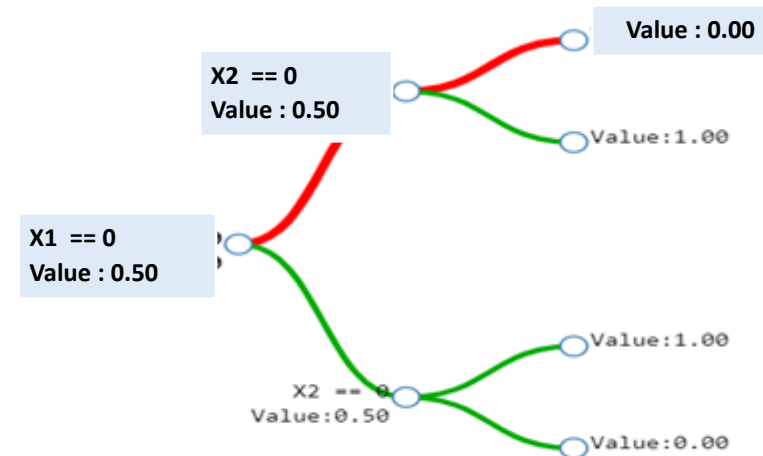
Conditional Feature Contribution From Tree Interpreter

There might be the case where features individually contributes nothing towards prediction but becomes predictive in conjunction with the other input feature

Exclusive OR(XOR)

| X1 | X2 | OUTPUT |
|----|----|--------|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

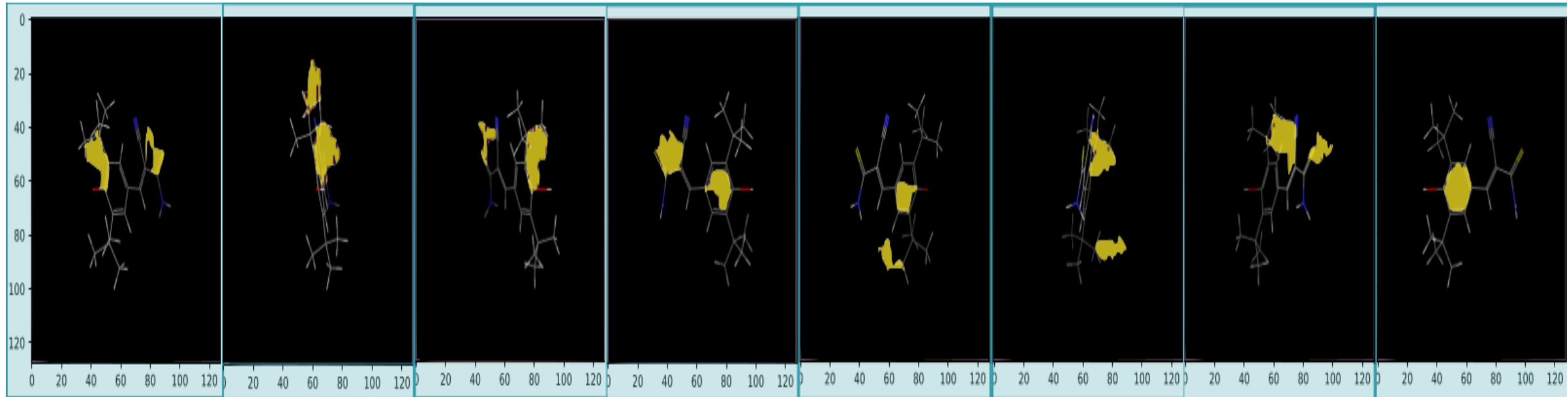
prediction = bias + feature₁contribution + feature₂contribution + (feature₁,feature₂)contribution



Prediction: 0.00 \approx 0.50 (trainset mean) + 0.00 (increase from X1) - 0.5 (decrease from X1,X2)

Results On Images

- A toxic compound which is predicted as toxic with a probability of 0.84 using our models and predicted as toxic with LIME's local linear model as toxic with a probability of 0.87. The most important two segments are shown.



Independent to the rotation our local model picked same region for four out of eight rotations.

Main Issues : No comparable measure across different images.

Results On Fingerprint Data

- To verify our approach we generated an ideal dataset to mimic our fingerprint dataset. In this ideal dataset we intentionally engineered the expected importance of the features.

| | feature0 | feature1 | feature2 | feature3 | feature4 | feature5 | feature6 | feature7 | feature8 | feature9 | ... | feature90 | feature91 | feature92 | feature93 |
|--------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----|-----------|-----------|-----------|-----------|
| perc_label_1 | 0.456 | 0.400 | 0.246 | 0.18 | 0.122 | 0.574 | 0.672 | 0.782 | 0.816 | 0.852 | ... | 0.536 | 0.524 | 0.492 | 0.498 |
| perc_label_0 | 0.524 | 0.656 | 0.754 | 0.79 | 0.860 | 0.448 | 0.374 | 0.274 | 0.192 | 0.156 | ... | 0.510 | 0.448 | 0.452 | 0.546 |

- We tested our model for few data points of ideal dataset and the output importance matched with the expected importance.

Test Case :

| | feature0 | feature1 | feature2 | feature3 | feature4 | feature5 | feature6 | feature7 | feature8 | feature9 | ... | feature91 | feature92 | feature93 |
|-----|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----|-----------|-----------|-----------|
| 600 | 0.0 | 0.0 | 1.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 0.0 | 0.0 |

Output :

| Feature Combination | Importance values |
|----------------------------------|-------------------|
| (Feature2 , Feature4, Feature9) | -0.4107 |
| (Feature4, Feature9) | -0.2614 |
| (Feature9 ,) | 0.2432 |

- ❖ For fingerprint dataset , since the deep learning model(black box) is trained on small data set the performance is not up to the mark with the approximate **balanced accuracy of 60%** and hence results could be unreliable.

Future Work

- Build a robust deep learning model for fingerprint dataset.
- Proper sampling across the local decision boundary.
- Develop a comparable measure to find out globally important features for all the images.
- We can check methods like LRP for further improvements in the model.

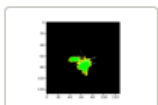
Thank You...



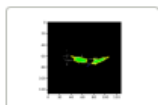
References

- [1] : <https://software.intel.com/en-us/articles/ai-helps-with-skin-cancer-screening>
- [2] : <https://searchhealthit.techtarget.com/feature/Use-of-AI-in-healthcare-seen-saving-lives-of-patients>
- [3,4,10,11] : Ece's CADD GRC presentation
- [5,6,9,12] : <https://github.com/marcotcr/lime>
- [13] : <https://www.oreilly.com/library/view/data-analysis-with/9781788393720/7d7c538e-2f7f-4e7e-9661-cd34f37c4711.xhtml>
- [14]: <http://blog.datadive.net/random-forest-interpretation-with-scikit-learn/>
- [15]: <https://blog.datadive.net/random-forest-interpretation-conditional-feature-contributions/>

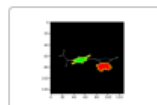
Back Up -1



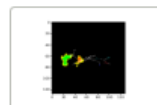
ID_NCGC00260731_angles_0_
45.pdf
7,7 kB



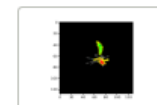
ID_NCGC00260731_angles_1_
45.pdf
8,1 kB



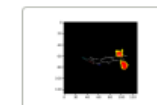
ID_NCGC00260731_angles_2_
45.pdf
8,4 kB



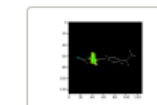
ID_NCGC00260731_angles_3_
45.pdf
8,1 kB



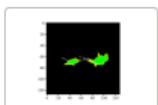
ID_NCGC00260731_angles_4_
45.pdf
7,7 kB



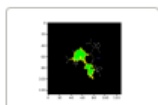
ID_NCGC00260731_angles_5_
45.pdf
8,2 kB



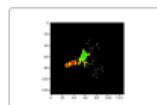
ID_NCGC00260731_angles_6_
45.pdf
8,3 kB



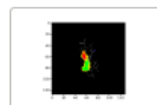
ID_NCGC00260731_angles_-
1_0.pdf
5,1 MB



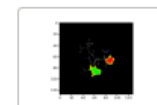
ID_NCGC00260864_angles_0_
45.pdf
9,4 kB



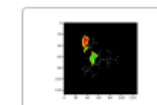
ID_NCGC00260864_angles_1_
45.pdf
9,3 kB



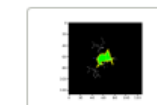
ID_NCGC00260864_angles_2_
45.pdf
8,9 kB



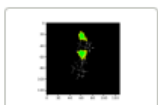
ID_NCGC00260864_angles_3_
45.pdf
9,1 kB



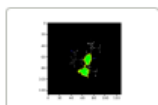
ID_NCGC00260864_angles_4_
45.pdf
9,7 kB



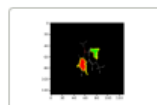
ID_NCGC00260864_angles_5_
45.pdf
9,1 kB



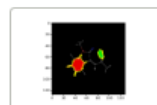
ID_NCGC00260864_angles_6_
45.pdf
9,1 kB



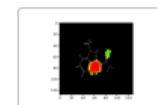
ID_NCGC00260864_angles_-
1_0.pdf
9,2 kB



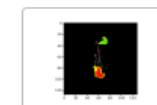
ID_NCGC00260922_angles_0_
45.pdf
8,7 kB



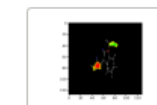
ID_NCGC00260922_angles_1_
45.pdf
9,3 kB



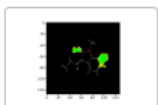
ID_NCGC00260922_angles_2_
45.pdf
9,4 kB



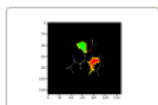
ID_NCGC00260922_angles_3_
45.pdf
8,4 kB



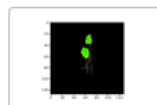
ID_NCGC00260922_angles_4_
45.pdf
8,8 kB



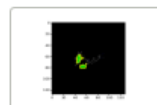
ID_NCGC00260922_angles_5_
45.pdf
9,6 kB



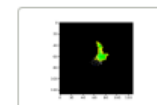
ID_NCGC00260922_angles_6_
45.pdf
9,2 kB



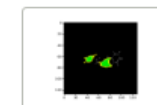
ID_NCGC00260922_angles_-
1_0.pdf
8,3 kB



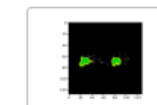
ID_NCGC00260954_angles_0_
45.pdf
7,7 kB



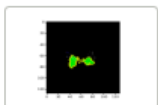
ID_NCGC00260954_angles_1_
45.pdf
7,5 kB



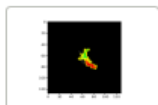
ID_NCGC00260954_angles_2_
45.pdf
8,0 kB



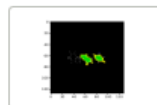
ID_NCGC00260954_angles_3_
45.pdf
8,2 kB



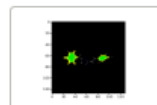
ID_NCGC00260954_angles_4_
45.pdf
7,8 kB



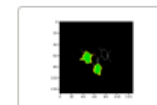
ID_NCGC00260954_angles_5_
45.pdf
7,3 kB



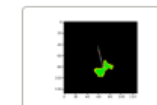
ID_NCGC00260954_angles_6_
45.pdf
7,9 kB



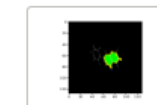
ID_NCGC00260954_angles_-
1_0.pdf
8,0 kB



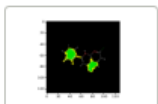
ID_NCGC00260990_angles_0_
45.pdf
8,7 kB



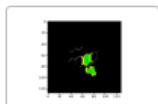
ID_NCGC00260990_angles_1_
45.pdf
7,4 kB



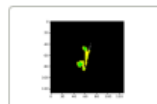
ID_NCGC00260990_angles_2_
45.pdf
8,7 kB



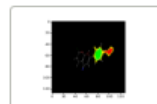
ID_NCGC00260990_angles_3_
45.pdf
9,0 kB



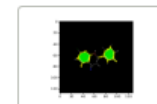
ID_NCGC00260990_angles_4_
45.pdf
8,7 kB



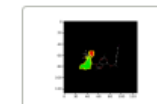
ID_NCGC00260990_angles_5_
45.pdf
7,0 kB



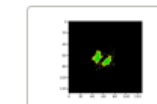
ID_NCGC00260990_angles_6_
45.pdf
8,8 kB



ID_NCGC00260990_angles_-
1_0.pdf
8,9 kB

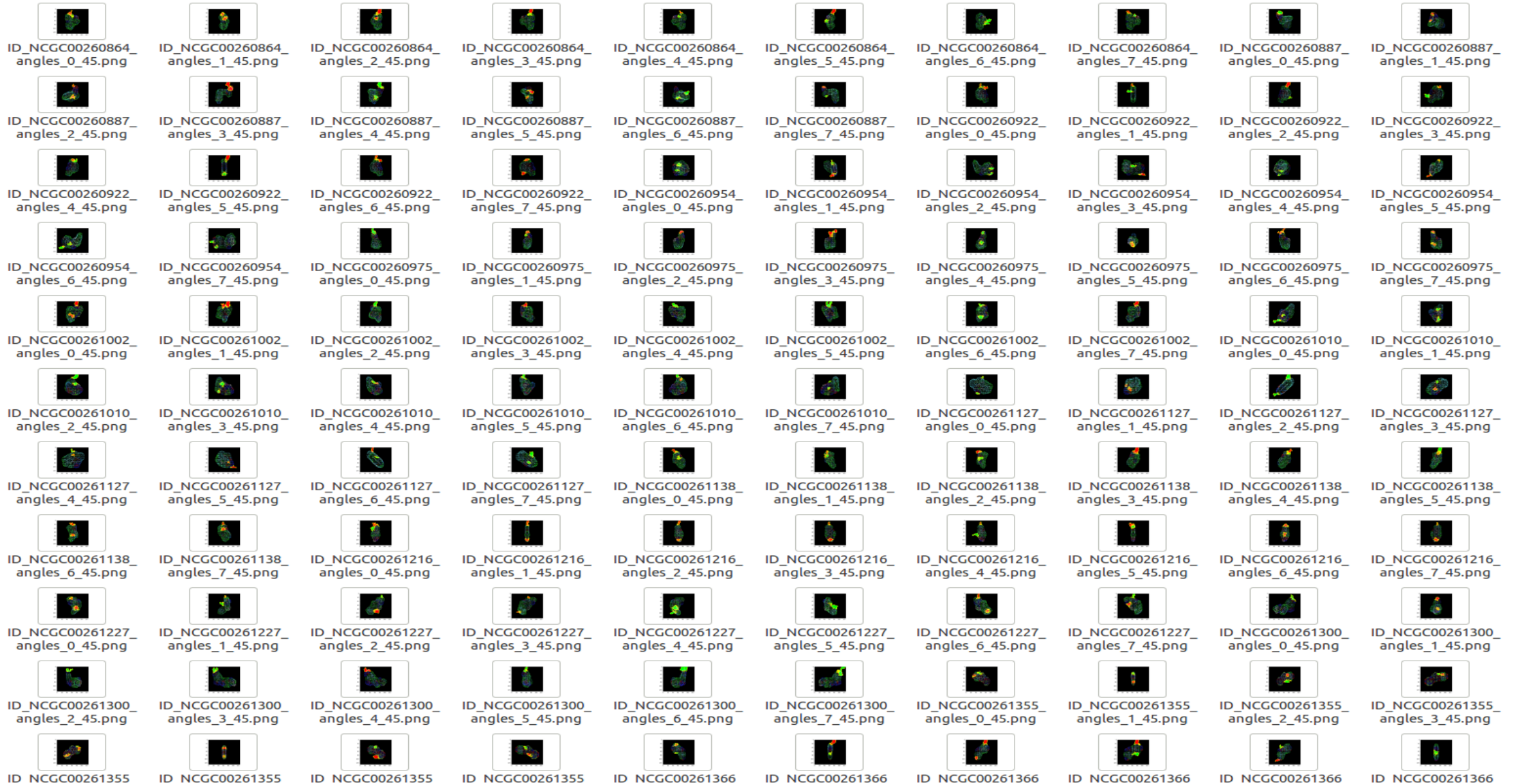


ID_NCGC00261010_angles_0_
45.pdf
9,2 kB



ID_NCGC00261010_angles_1_
45.pdf
9,0 kB

Back Up - 2



Back Up - 3

Ideal Dataset


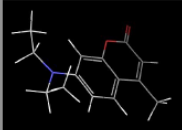
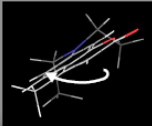
| | feature0 | feature1 | feature2 | feature3 | feature4 | feature5 | feature6 | feature7 | feature8 | feature9 | ... | feature91 | feature92 | feature93 |
|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----|-----------|-----------|-----------|
| 0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.0 | ... | 0.0 | 0.0 | 1.0 |
| 1 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | ... | 0.0 | 1.0 | 1.0 |
| 2 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | ... | 1.0 | 0.0 | 1.0 |
| 3 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 0.0 | 0.0 | 1.0 | ... | 0.0 | 1.0 | 0.0 |
| 4 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 | 1.0 | ... | 0.0 | 1.0 | 0.0 |

Description of the Datasets

Tox Alert dataset

| Toxic | Non-Toxic |
|-------|-----------|
| 611 | 497 |

Tox21 p53 DataSet

| | | Toxic | Not Toxic |
|---|----------------|-------|-----------|
|  | 2D | 371 | 5741 |
|  | 3D | 10865 | 11001 |
|  | After Rotation | 86929 | 88017 |