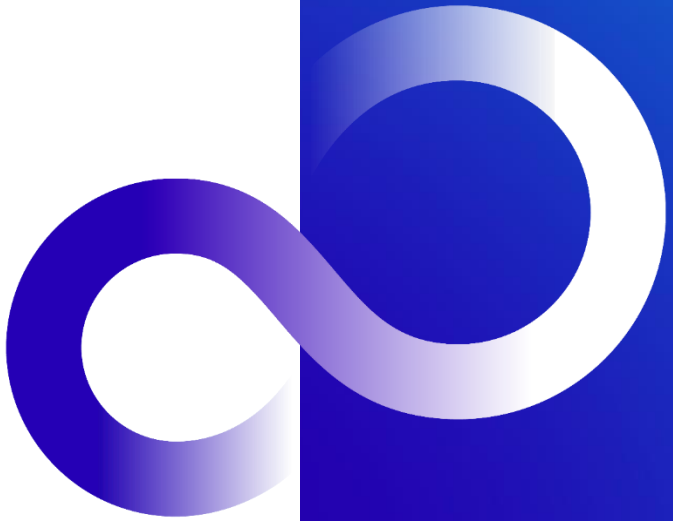
Machine Learning Models for fast estimation of *electronic and thermodynamic* properties of small molecules

Wojtek Plonka
FQS Poland
w.plonka@fqs.pl



Why replace DFT with ML?

What is QSAR?

FUJITSU

International Journal of Quantitative Structure-Property Relationships
Volume 1 • Issue 1 • January-June 2016

The History and Development of Quantitative Structure-Activity Relationships (QSARs)

John C. Dearden, School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Liverpool, UK

ABSTRACT

It is widely accepted that modern QSAR began in the early 1960s. However, as long ago as 1816 scientists were making predictions about physical and chemical properties. The first investigations into the correlation of biological activities with physicochemical properties such as molecular weight and aqueous solubility began in 1841, almost 60 years before the important work of Overton and Meyer linking aquatic toxicity to lipid-water partitioning. Throughout the 20th century QSAR progressed, though there were many lean years. In 1962 came the seminal work of Corwin Hansch and co-workers, which stimulated a huge interest in the prediction of biological activities. Initially that interest lay largely within medicinal chemistry and drug design, but in the 1970s and 1980s, with increasing ecotoxicological concerns, QSAR modelling of environmental toxicities began to grow, especially once regulatory authorities became involved. Since then QSAR has continued to expand, with over 1400 publications annually from 2011 onwards.

Why replace DFT with ML?

- Because it *should* be much faster to compute
- To test our ML methodology on reasonable data
- Why not?

Materials and Methods

The data – QM9

Properties calculated by B3LYP

https://moleculenet.org/datasets-1
Molecules with extreme values rejected – top and bottom 1%

These molecules correspond to the subset of all 133,885 species with up to nine heavy atoms (CONF) out of the GDB-17 chemical universe of 166 billion organic molecules.

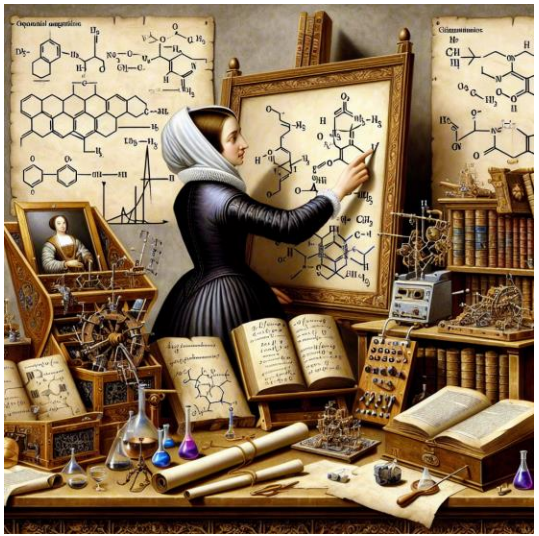
R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, Scientific Data 1, 140022, 2014.

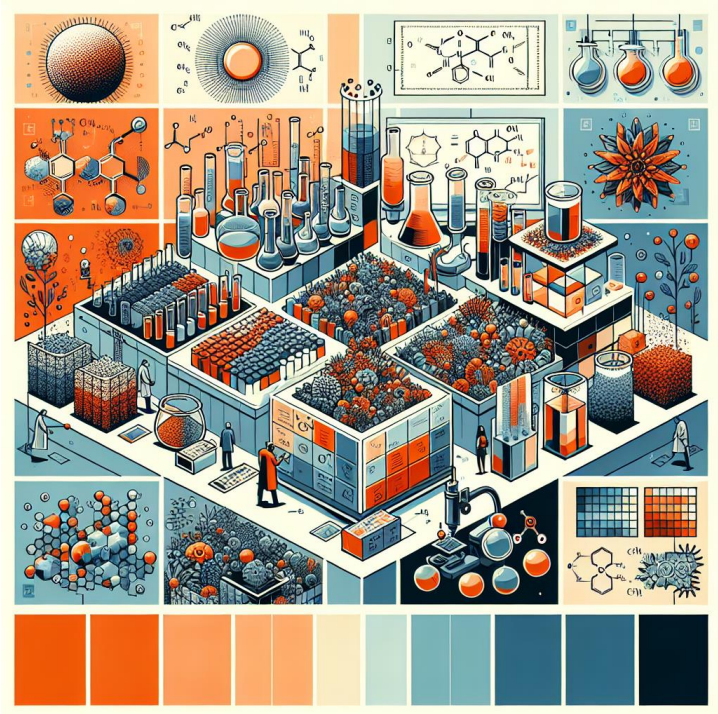
μ	D	Dipole moment
α	a_0^3	Isotropic polarizability
ϵ_{HOMO}	Ha	Energy of HOMO
ϵ_{LUMO}	Ha	Energy of LUMO
ϵ_{gap}	Ha	Gap ($\epsilon_{LUMO} - \epsilon_{HOMO}$)
$\langle R^2 \rangle$	a_0^2	Electronic spatial extent
zpve	Ha	Zero point vibrational energy
U_o	Ha	Internal energy at 0 K
U	Ha	Internal energy at 298.15 K
H	Ha	Enthalpy at 298.15 K
G	Ha	Free energy at 298.15 K
C_v	$\frac{\text{cal}}{\text{molK}}$	Heat capacity at 298.15 K

The steps to build QSAR model

FUJITSU

- Convert chemistry to mathematics
- Train and validate the model





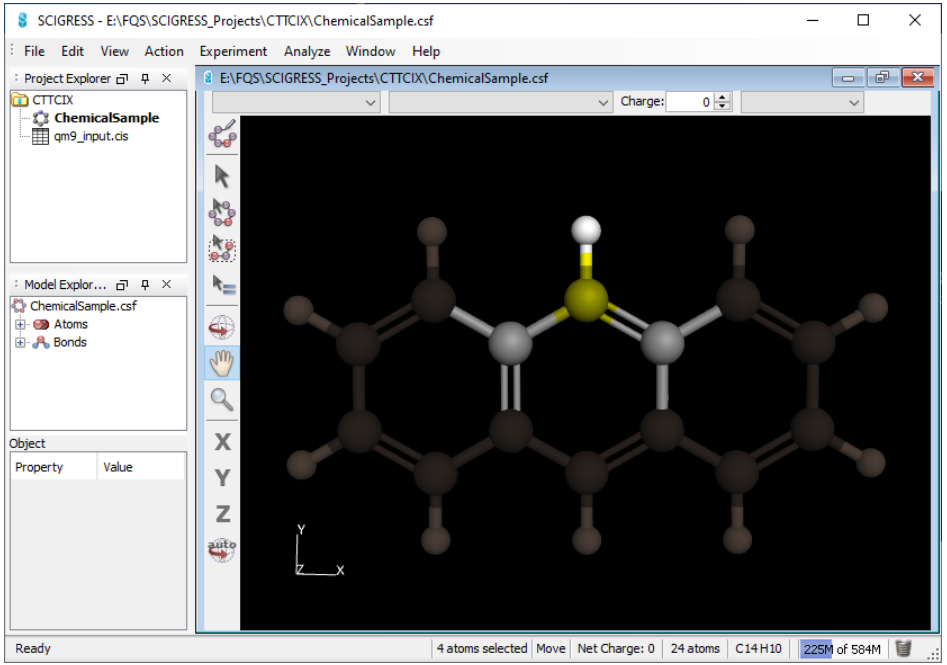
Convert chemistry to mathematics

FUJITSU

- Use Morgan fingerprint descriptors
  - They are fast
  - They are robust
  - They are well-proven

https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

Morgan Fingerprint is...

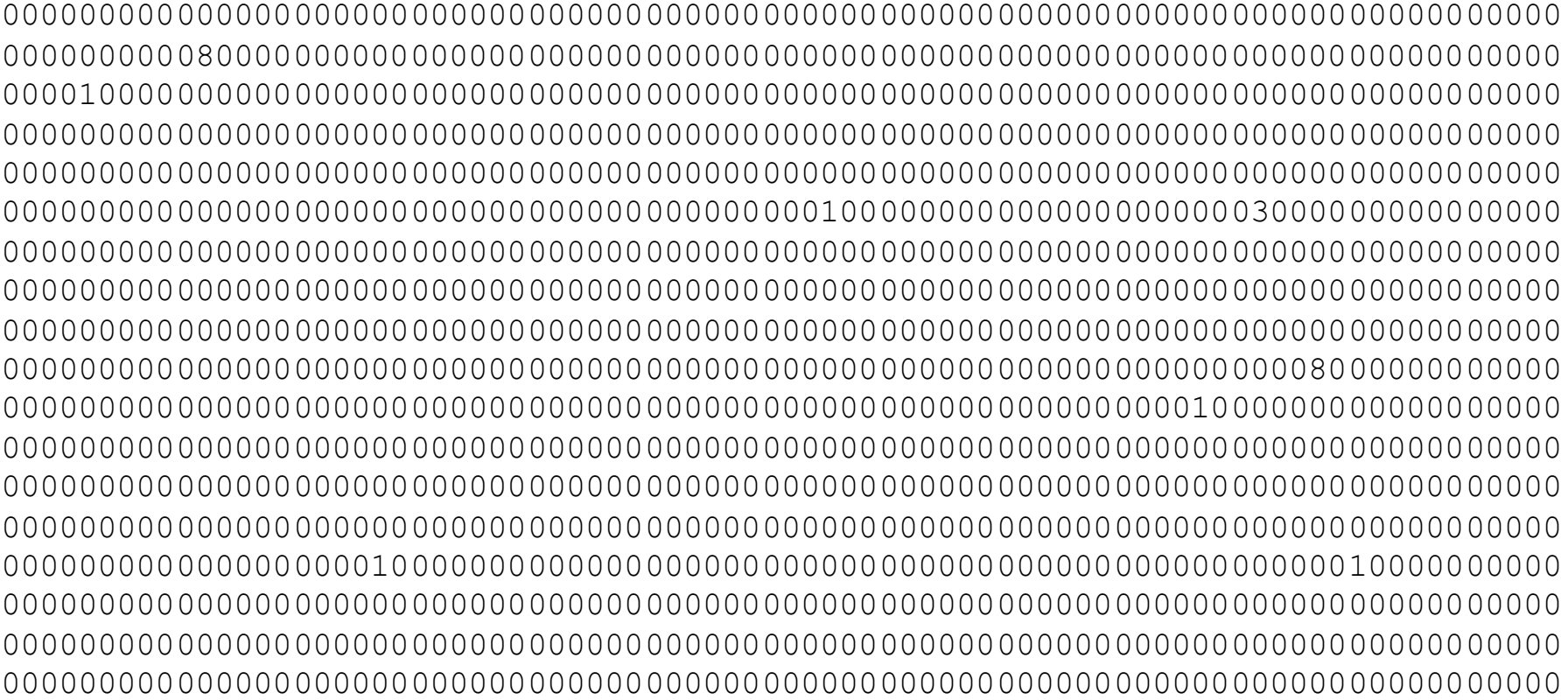



```
mol = Chem.MolFromSmiles('c1ccc2cc3ccccc3cc2c1')
info = {}
fp = AllChem.GetMorganFingerprint(mol, radius = 1,
fp = fp.GetNonzeroElements()
fp
```

[11] ✓ 0.0s

```
{98513984: 4,
 951226070: 4,
 994485099: 2,
 2360741695: 4,
 3217380708: 4,
 3218693969: 10}
```

...a lot of empty space



Hashing - bits

00000010000001000010000000100000000100000

0000001
0000001
0000100
0000010
0000000
0100000

0100111

Hashing - counts

FUJITSU

00000010000007000020000001000000000500000

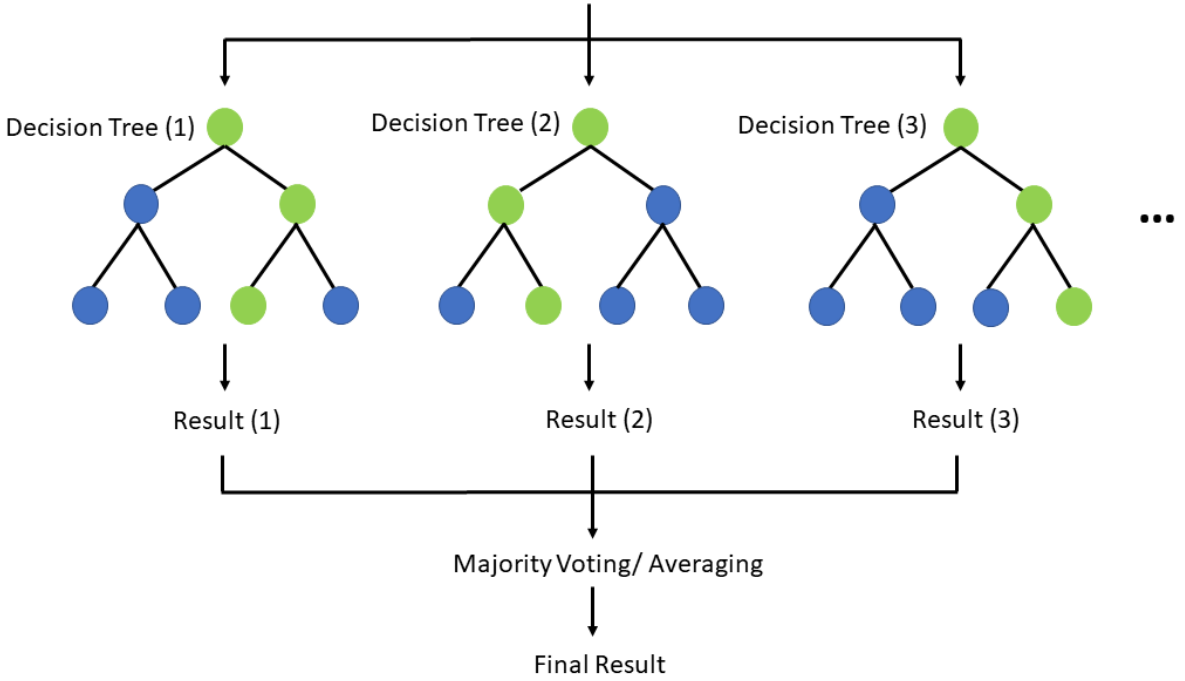
0000001
0000007
0000200
0000010
0000000
0500000

0500218

Time to learn...

FUJITSU

Random Forest Regressor



TseKiChun, CC BY-SA 4.0 <https://creativecommons.org/licenses/by-sa/4.0>, via Wikimedia Commons

...and to judge!

FUJITSU

Internal Split K-Fold Cross-Validation Using R^2 in Regression involves dividing the dataset into *k* subsets, or "folds," and then performing multiple iterations of training and validation to ensure that the model generalizes well to unseen data.

Steps in K-Fold Cross-Validation

1. **Data Splitting**: The dataset is randomly divided into *k* equal-sized folds.
2. **Model Training and Validation**: For each fold, the model is trained on *k*-1 folds and validated on the remaining fold. This process is repeated *k* times, with each fold serving as the validation set once.
3. **Performance Evaluation**: The performance metric, such as R^2 (coefficient of determination), is calculated for each iteration. The overall performance is typically assessed by averaging the metric across all folds.

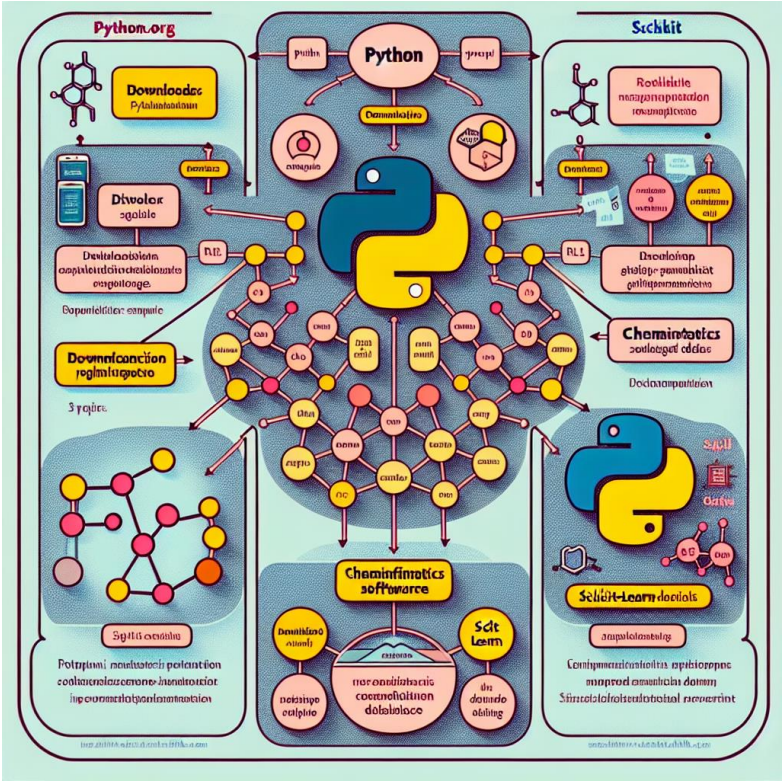
R^2 in Regression

R^2 is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. It provides an indication of how well the model predictions fit the actual data. In the context of k-fold cross-validation, R^2 is computed for each fold, and the average R^2 across all folds is used as an indicator of the model's performance.

perplexity.ai

The toolbox

- python.org
- rdkit.org
- scikit-learn.org



Results

The winners and the losers

FUJITSU

General Guidelines for R^2 Values

Physical Sciences: In fields like physics or chemistry, where processes are often well-understood and measurements are precise, R^2 values are typically expected to be high, often above **0.9**. This reflects a strong relationship between the variables and a high degree of predictability

Social Sciences: In social sciences, such as psychology or sociology, the data often involve complex human behaviors that are harder to predict. Here, lower R^2 values are more common, and an R^2 as low as **0.1** can be considered acceptable if the predictors are statistically significant. This is because the focus is often on understanding the impact of specific variables rather than achieving high predictive accuracy.

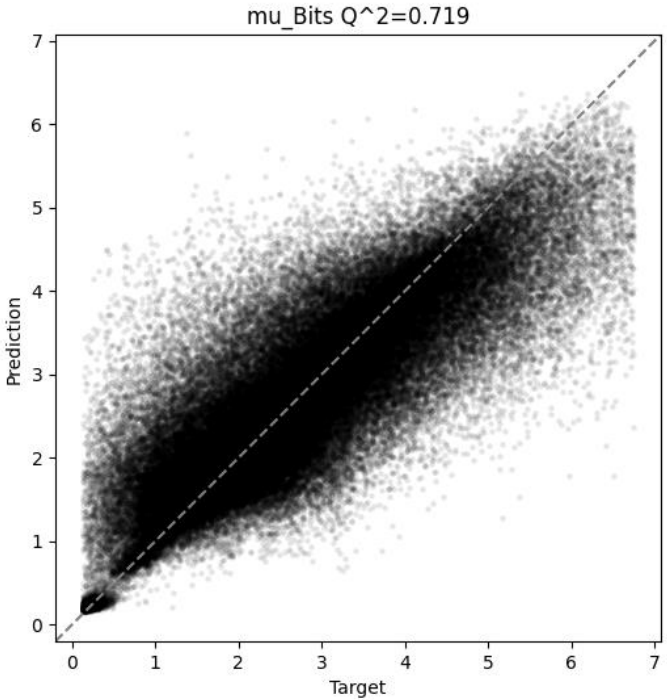
Life Sciences: In fields like biology or ecology, the acceptable R^2 threshold can vary widely. For some studies, especially those involving complex biological systems, R^2 values might be lower due to inherent variability and complexity

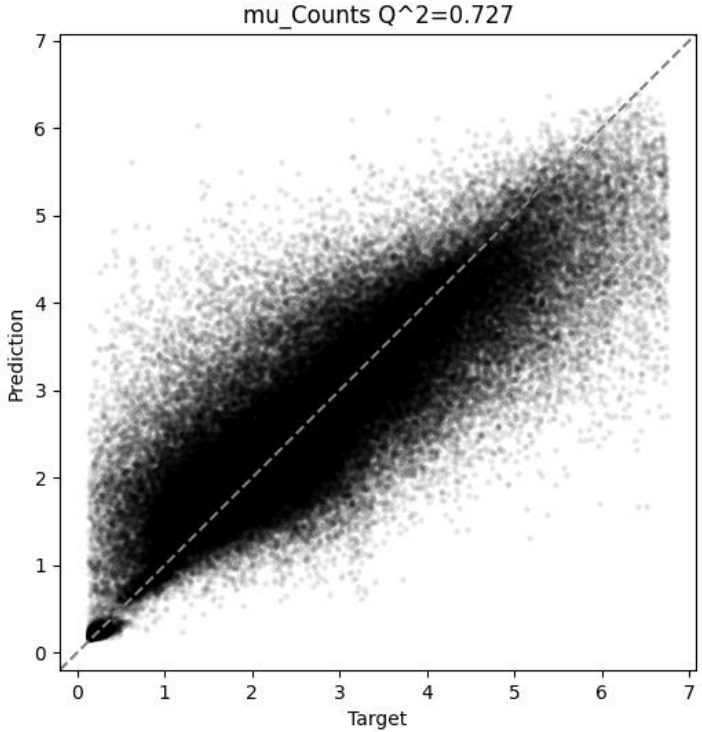
perplexity.ai

The summary

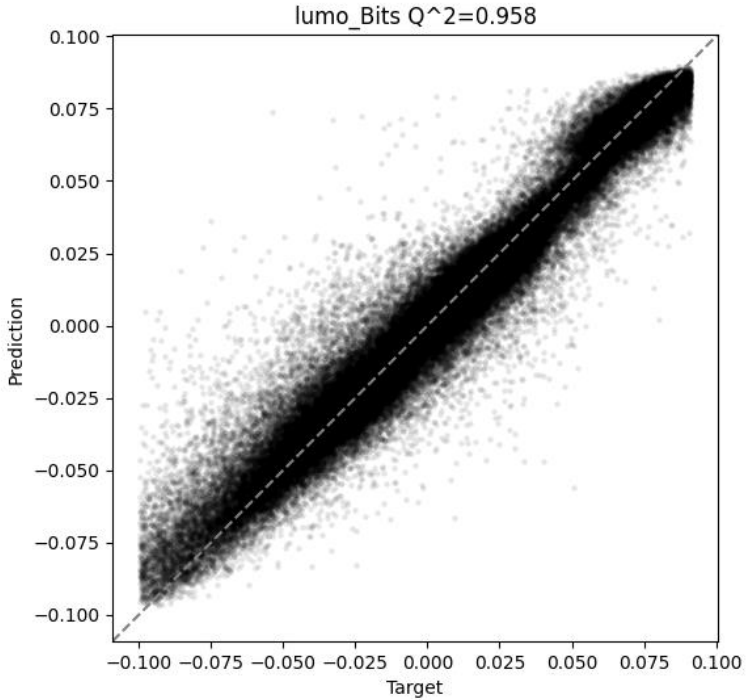
	R^2 bits	R^2 counts
Dipole moment	0.719	0.727
Isotropic polarizability	0.606	0.804
Energy of HOMO	0.839	0.865
Energy of LUMO	**0.958**	**0.968**
HOMO - LUMO gap	0.948	**0.956**
Electronic spatial extent	0.742	0.816
Zero point vibrational energy	0.857	**0.971**
Internal energy at 0K	0.682	0.813
Internal energy at 298.15K	0.682	0.813
Enthalpy at 298.15K	0.683	0.813
Free energy at 298.15K	0.682	0.813
Heat capacity at 298.15K	0.740	0.873

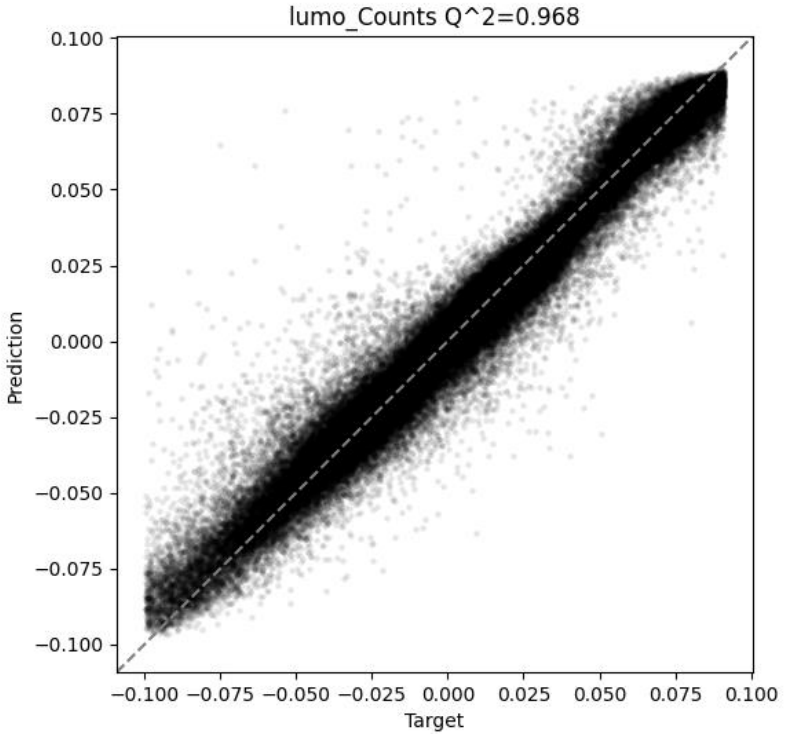
Dipole Moment



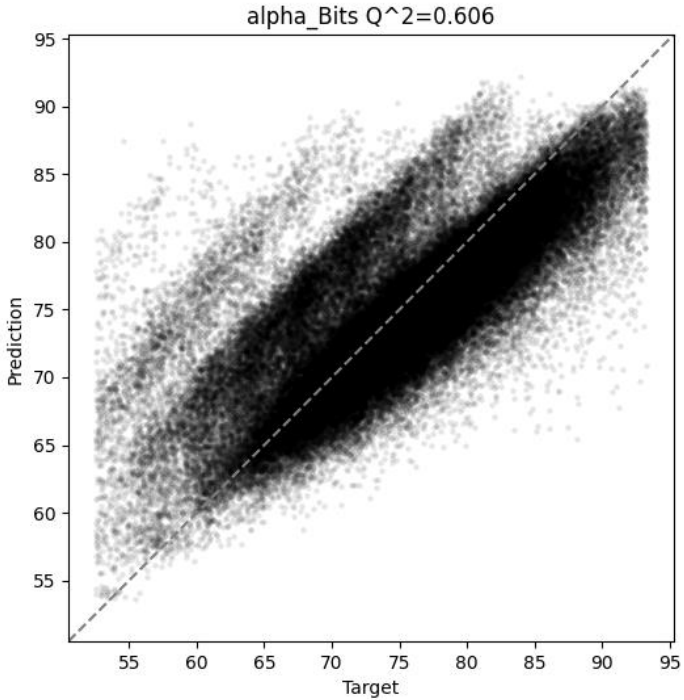


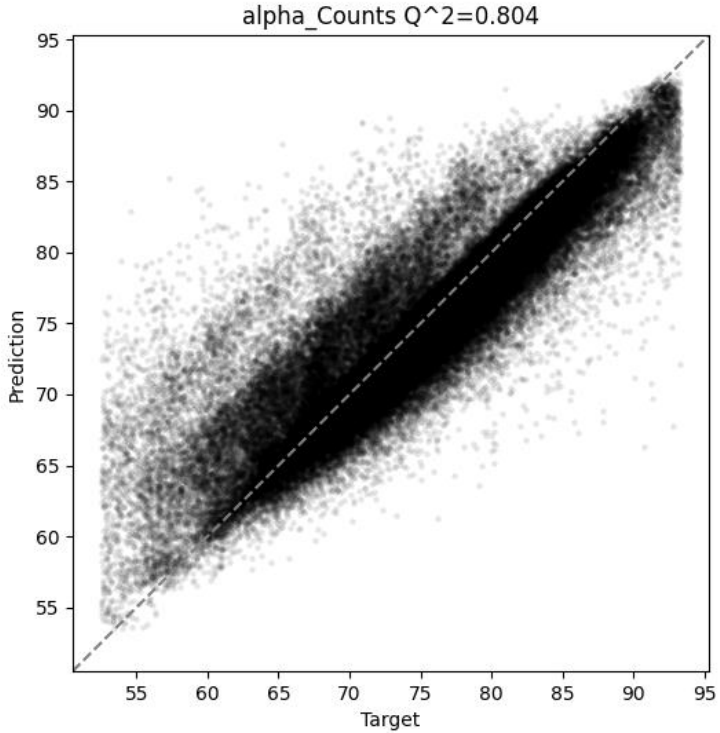
LUMO Energy



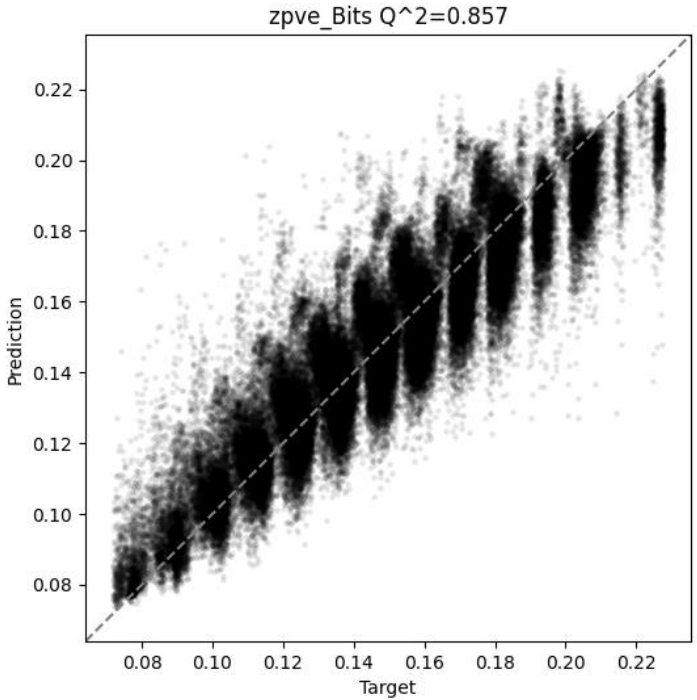


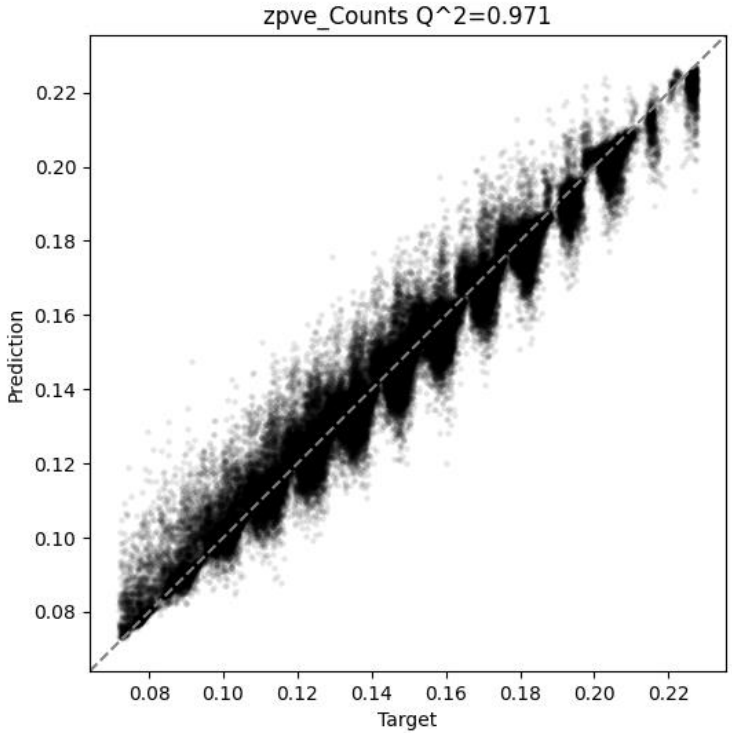
Isotropic polarizability





Zero point vibrational energy





FUJITSU

Conclusions?

Is ML worth it?

FUJITSU

- It works reasonably, for most cases.
- Counts of fingerprints are better than bits!
- There still is room for improvement… I hope.
- It's FAST!
  - 15 minutes to compute all the descriptors for whole data set
  - All models were created and evaluated overnight on desktop
- It's simple.

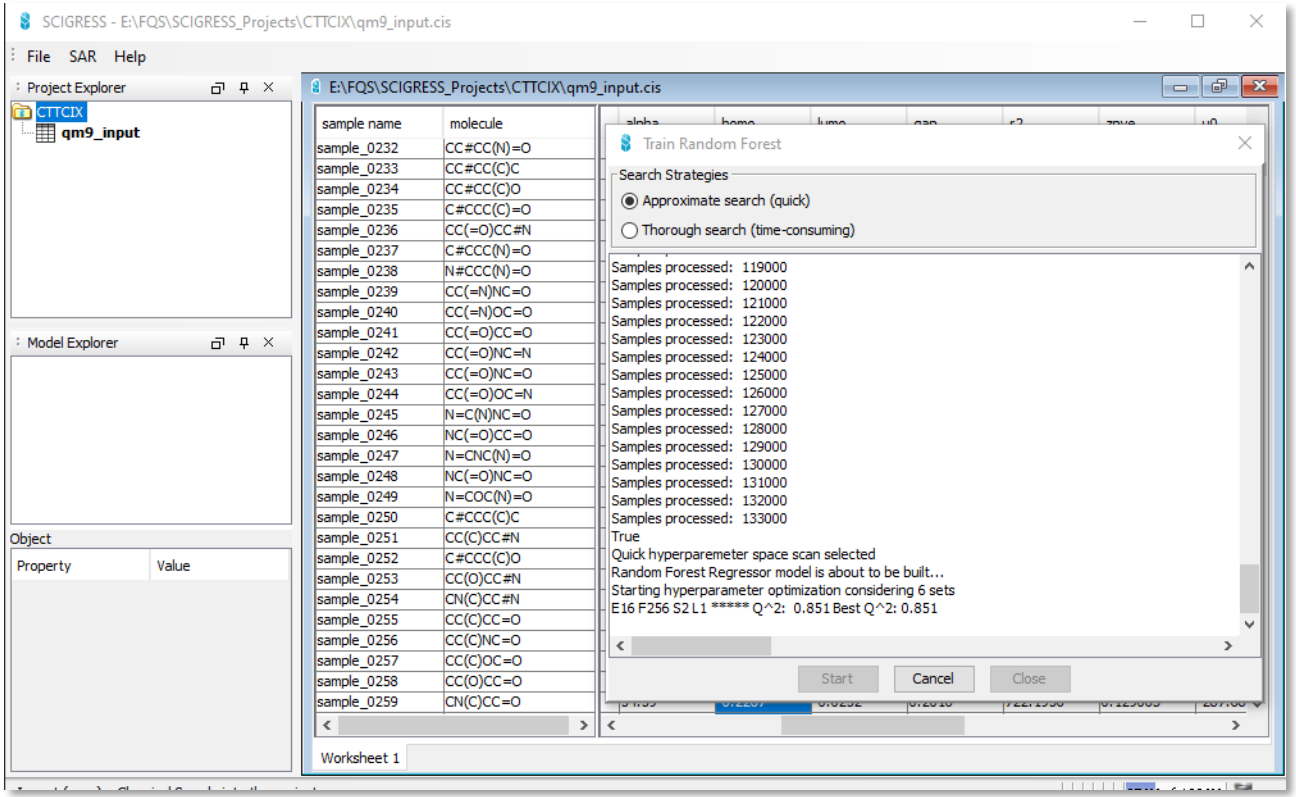
```
116         builModelV1R()
117 end_time = time.time()
118 elapsed_time = end_time - start_time
119 print(f"Elapsed time: {elapsed_time} seconds")
```

The easy, lazy way...

FUJITSU

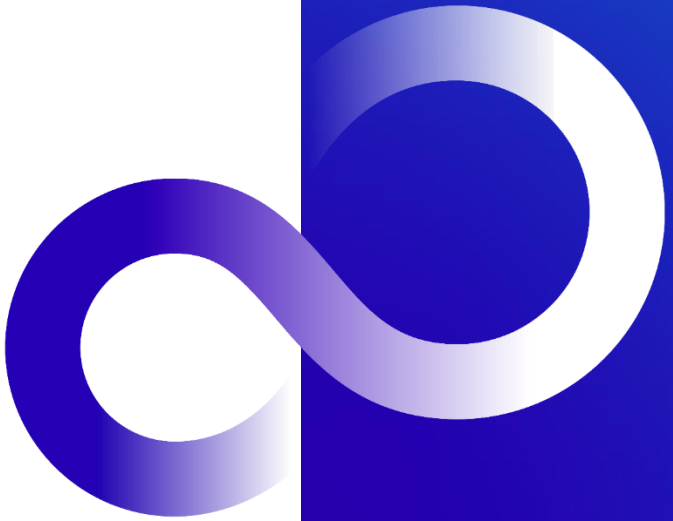
SCIGRESS

www.scigress.com



Thank you!

chemistry@fqs.pl



Molecular Weight