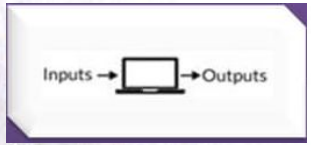
- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

Machine Learning

Machine learning techniques can be divided into two foremost types:

- **Unsupervised**: find hidden patterns or intrinsic structures in data. They are used to draw inferences from data sets consisting of input data without labeled responses (i.e. clustering algorithms)
- **Supervised**: used when you want to predict or explain the data you possess. A supervised algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions
- **Reinforcement Learning**: the algorithms learn to react to an environment on their own. An agent is in a situation of trial and error, where the consequences of its actions have an impact on the environment and also on the problem's goal. The agent is punished or rewarded on the basis of its behavior, with the idea that, in the future, it will prefer optimal actions (i.e. our intelligent cache system)



Tommaso Tedeschi, Marco Baioletti, Diego Ciangottini, Valentina Poggioni, Daniele Spiga, Loriano Storchi, Mirco Tracolli, "Smart Caching in a Data Lake for High Energy Physics Analysis", Journal of Grid Computing, DOI: 10.1007/s10723-023-09664-z (2023)

Machine Learning

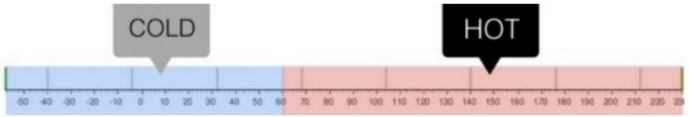
Regression

What will be the temperature tomorrow?

Fahrenheit

Classification

Will it be hot or cold tomorrow?



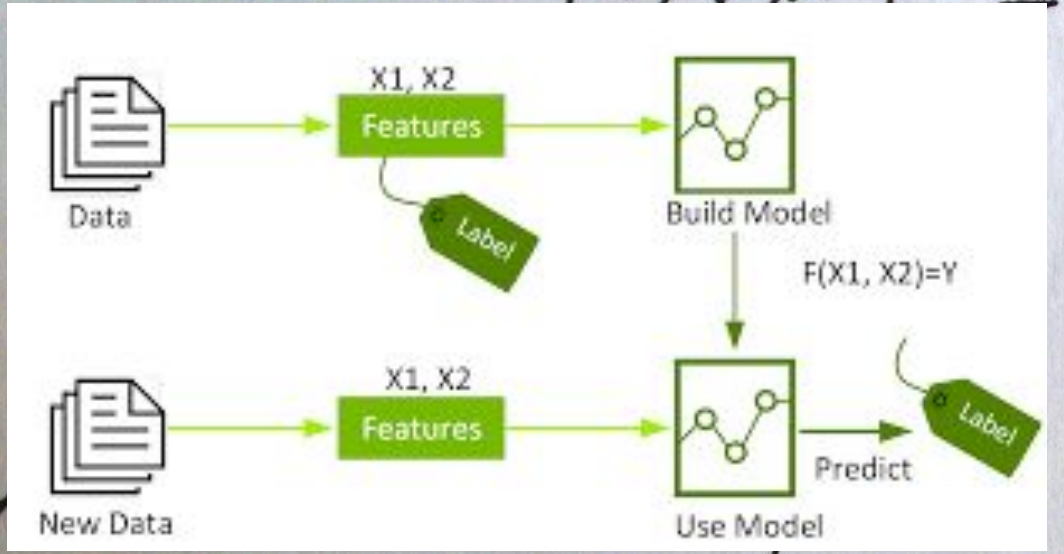
Fahrenheit

Features could be: the day of the year and the today temperature

Label: is the temperature for the regression and

Machine Learning

Supervised: used when you want to predict or explain the data you possess. A supervised algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions



$Y = F_{abc}(X)$

Labels: dependent variables (e.g. pK_a values , could be also a class pass or not the BBB)

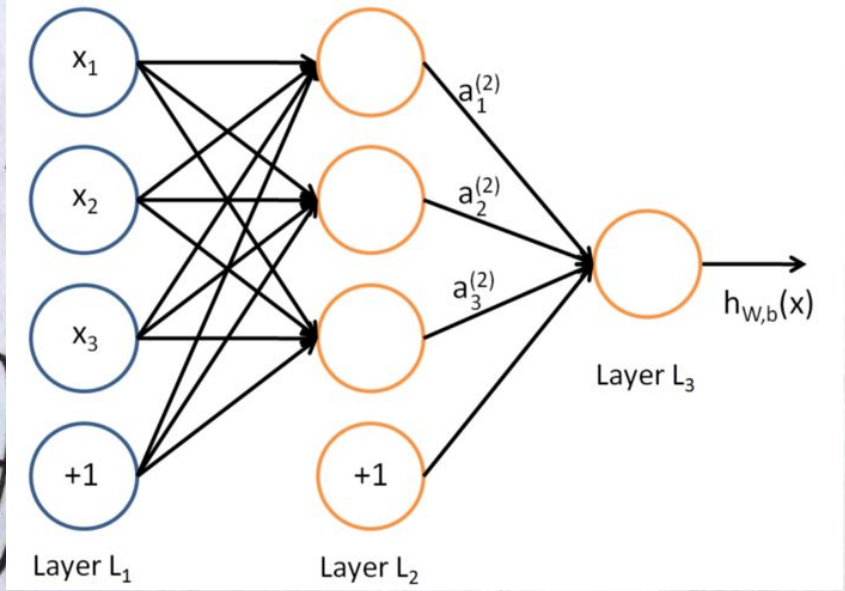
Features (descriptors): independent variables (e.g. Molecular weight, fingerprints)

Models: Linear Regression, Random Forest, Artificial Neural Network , Partial Leat Square

- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

Neural Network

- A layer is a collection of neurons which take an input and provide an output
- If there is more than 1 hidden layer then it is called a **Deep Neural Network**



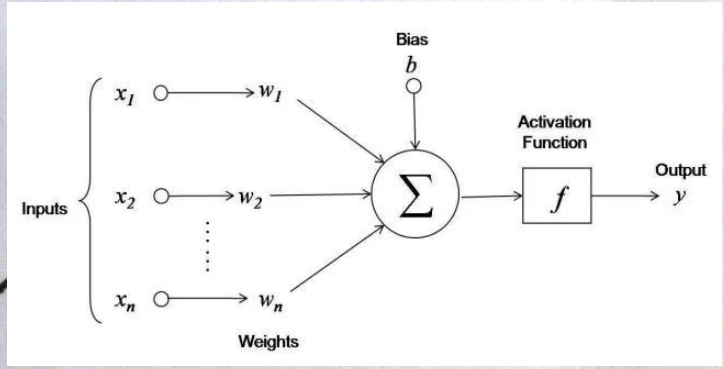
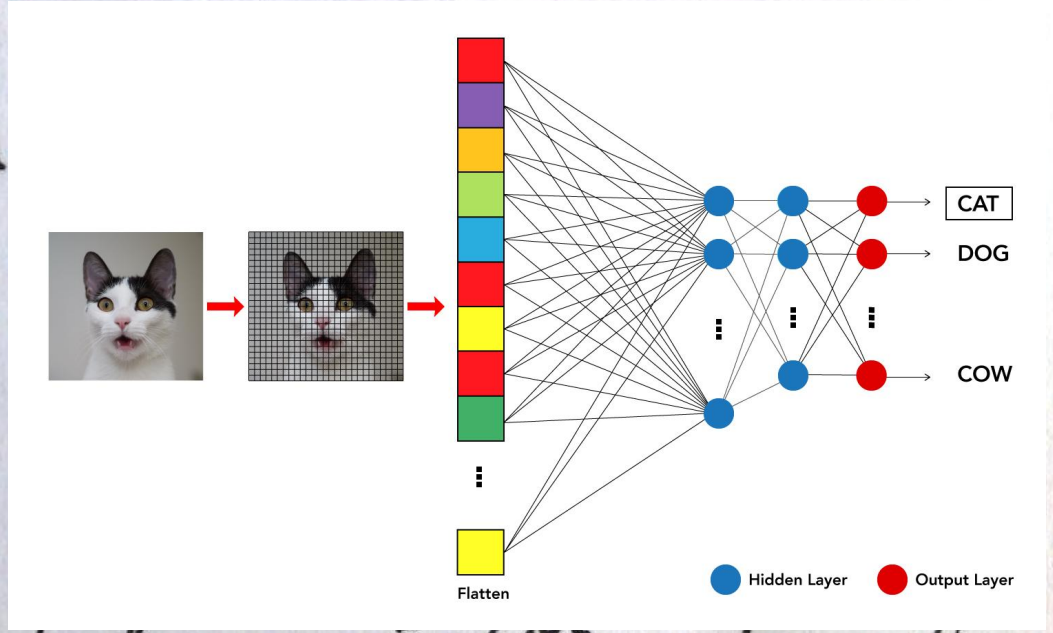


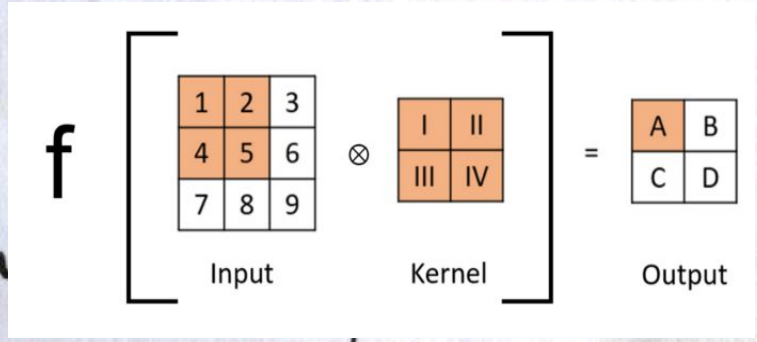
Image Recognition

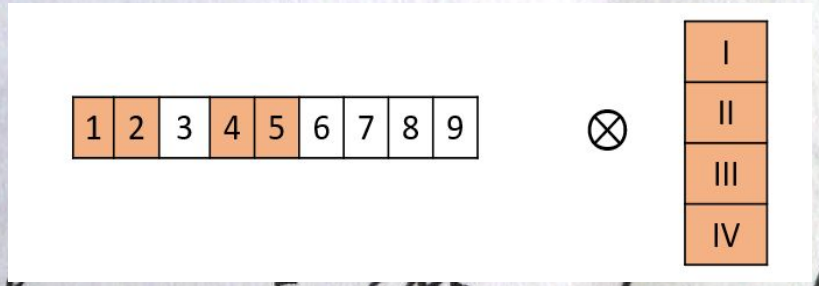
- Recognition of people, animals, objects, places etc from digital images
- Trained using thousands of pre-labelled images
- Uses the pixels in each image as descriptors
- Trained to recognise if the image shows a certain class



Convolutional Layers – extracting feature

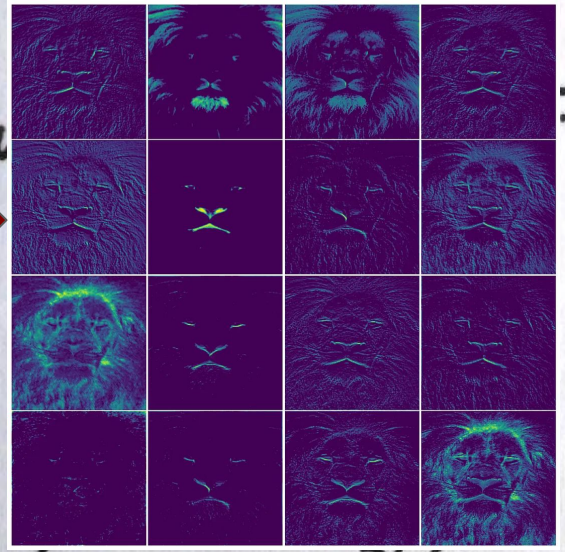
- An image is a cuboid having its length, width (dimension of the image), and height (i.e the channel 3 channelrs for RGB
- Kernel slides across the height and width of the image input and dot product of the kernel and the image are computed





Convolutional Layers – extracting feature

Convolutional layers often detect edges and geometries in the image (Colors: RGB three channels)



Predicting Gene Accessibility using CNNs

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*. 2016;26(7):990-999. doi:10.1101/gr.200535.115.

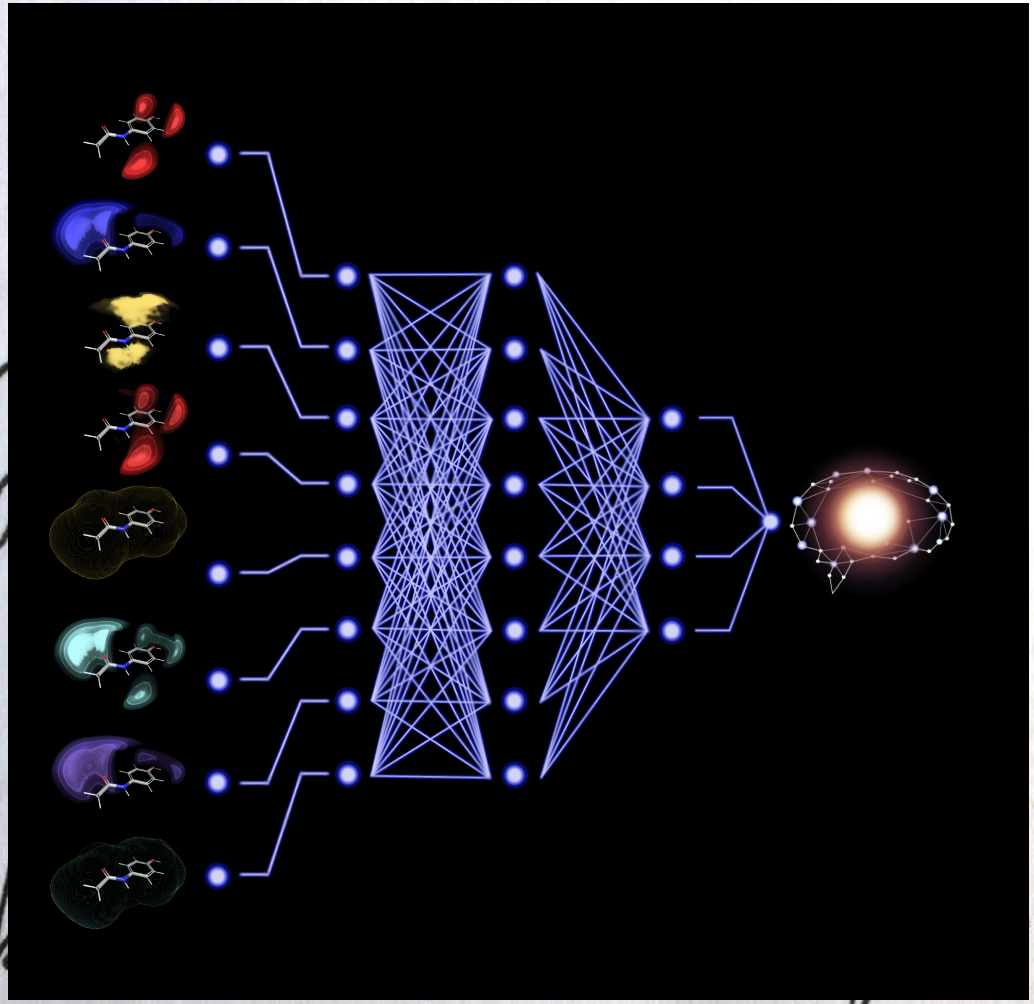
- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

DeepGRID

Two ingredients are needed:

- Deep Learning techniques (i.e., CNN)
- GRID MIFs

Loriano Storchi, Gabriele Cruciani, Simon Cross, "DeepGRID: Deep Learning using GRID descriptors for BBB prediction", Journal of Chemical Information and Modeling, DOI: 10.1021/acs.jcim.3c00768 (2023).



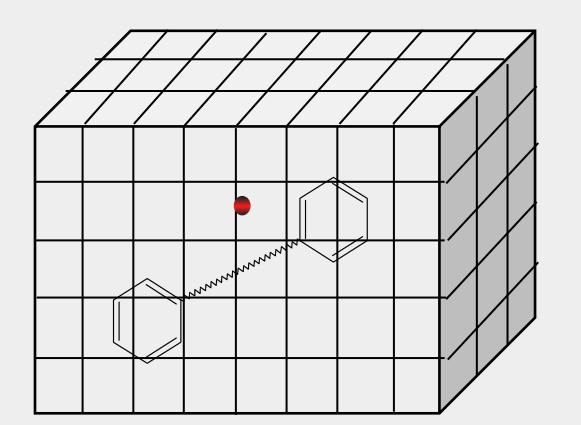
- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

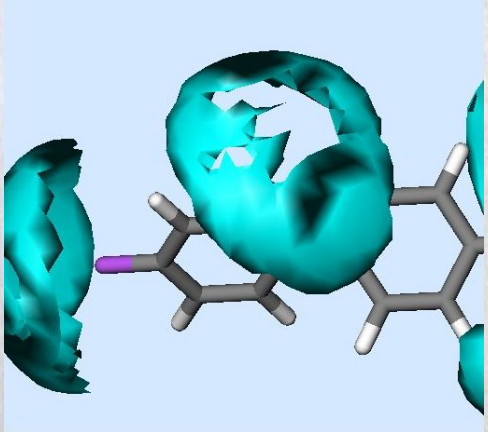
Machine Learning and the GRID Force-Fields

- **GRID program**: a computational procedure for determining energetically favourable binding sites on molecules for functional groups of known structure through the use of PROBES.
  - The PROBE is moved through a grid of points superimposed on the target molecule (to each atoms of the target and AtomType is assigned) . Its interaction energy with the target molecule is computed by an empirical energy function

$$E_{XYZ} = \Sigma[E_{LJ}] + \Sigma[E_{HB}] + \Sigma[E_{Q}] + [S]$$

E_{LJ}= Lennard-Jones potential E_{HB}= hydrogen bonding interaction energy E_{Q}= electrostatic function S= entropic term





Machine Learning and the GRID Force-Fields

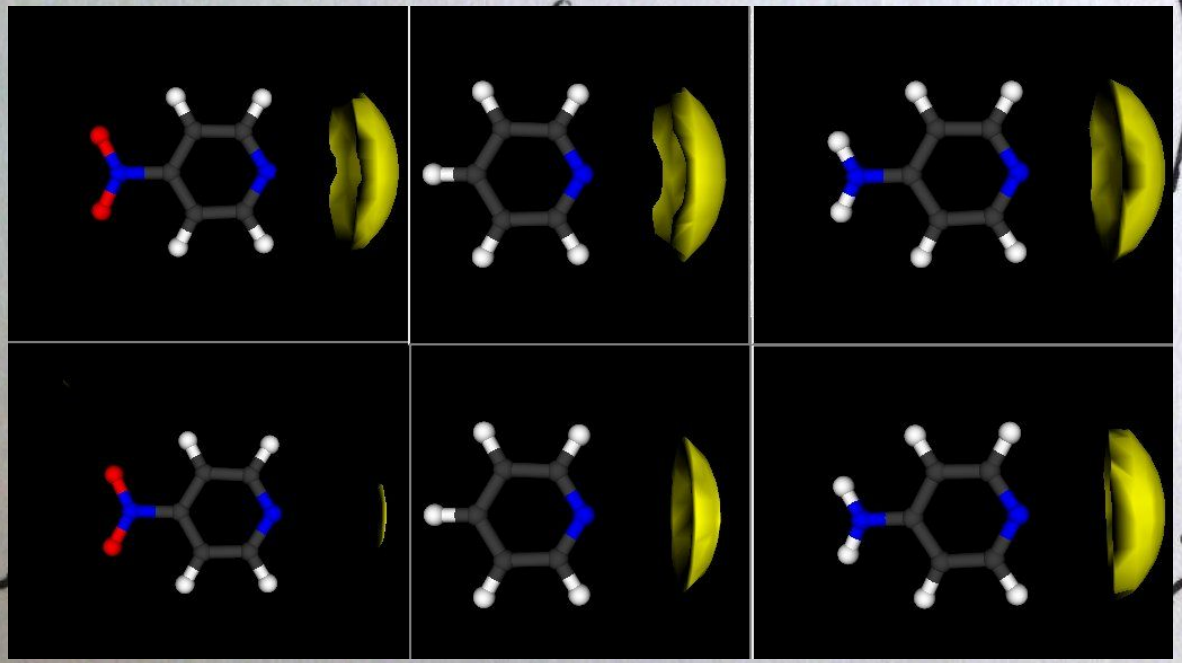
We build PLS models, each model is related to a specific AT, to improve the quality of the Hydrogen-Bonding term E_{HB}

$$E_{HB} = E_r * E_t * E_p.$$

Sara Tortorella, Emanuele Carosati, Giovanni Bocci, Simon Cross, Gabriele Cruciani, Loriano Storchi, "Combining Machine Learning and Quantum Mechanics Yields More Chemically-Aware Molecular Descriptors for Medicinal Chemistry Applications", Journal of Computational Chemistry, DOI: 10.1002/jcc.26737 (2021)

Machine Learning and the GRID Force-Fields

More chemically aware force-field



The energy values of the isocontour surfaces chosen for H-bond donating probe (“N1,” probe) was 4.0 kcal/Mol

- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

t's try it anyway...

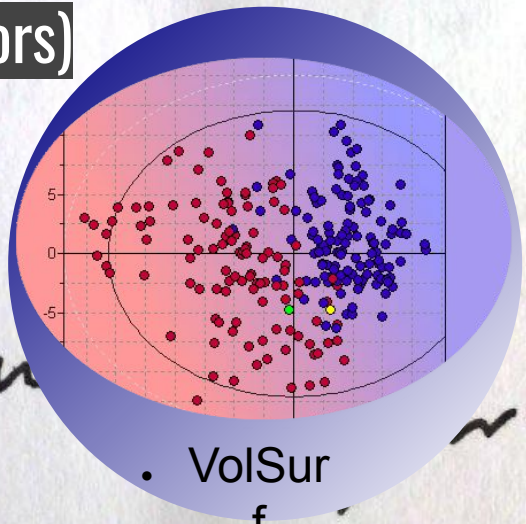
$$(i\gamma^\mu \partial_\mu - m)\psi = 0$$

DATASET AND LABEL

motion, want this, but $i\gamma^\mu$ not four

Test Case: Blood Brain Barrier Permeation

- A model exists within VolSurf (PLS) – we have a baseline
- We can investigate a number of modelling approaches: DeepGRID, Random Forest & PLS (using VS descriptors)
- There are some larger publicly available datasets eg. LightBBB (7000 cpds)



Dataset Preparation

- **VS-IgBB-332** dataset In-house dataset used to build the original VolSurf model
- **Light-BBclass-2105** dataset - Classification Generated from the Shaker/Parakkal LightBBB dataset of 7000+ structures
  - After filtering by InChI to remove duplicates 4285 compounds remained (-40%!)
  - Given that such a large proportion of the dataset contained duplicates we filtered also by Druglikeness to give 3464 compounds
  - 70% of the dataset removed due to duplicate InChI strings or diastereoisomerism
- **Light-IgBB-416** dataset A subset of the 2105 dataset which had experimental logBB values

Dataset Splitting

- For each dataset, subsets of compounds were randomly selected:
  - Training Set: 60% - used to train the models
  - Validation Set: 20% - used to select the best hyperparameters or to train the CNN
  - Test Set: 20% - used as a final performance check
- The same sets were used for each model

t's try it anyway...

$$(i\gamma^\mu \partial_\mu - m)\psi = 0$$

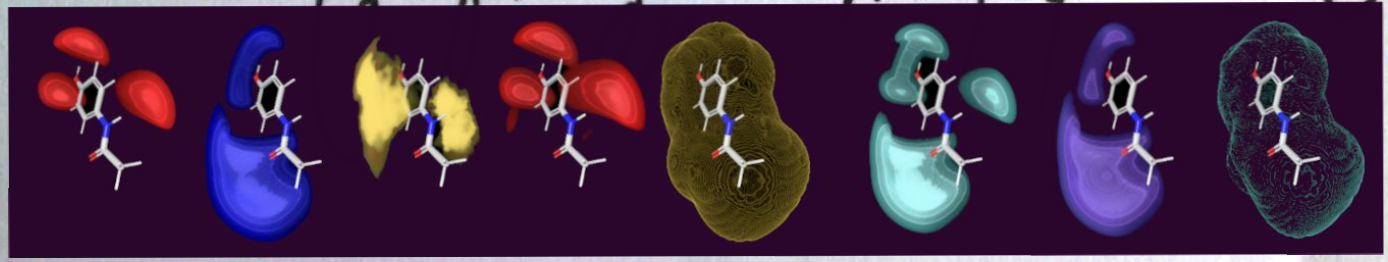
FEATURES

motion, want this, but $i\gamma^\mu$ not four

DeepGRID Approach

GRAID descriptors calculated (normalised GRID MIFs, 8 channels)

Descriptors fed into a Deep Learning CNN model

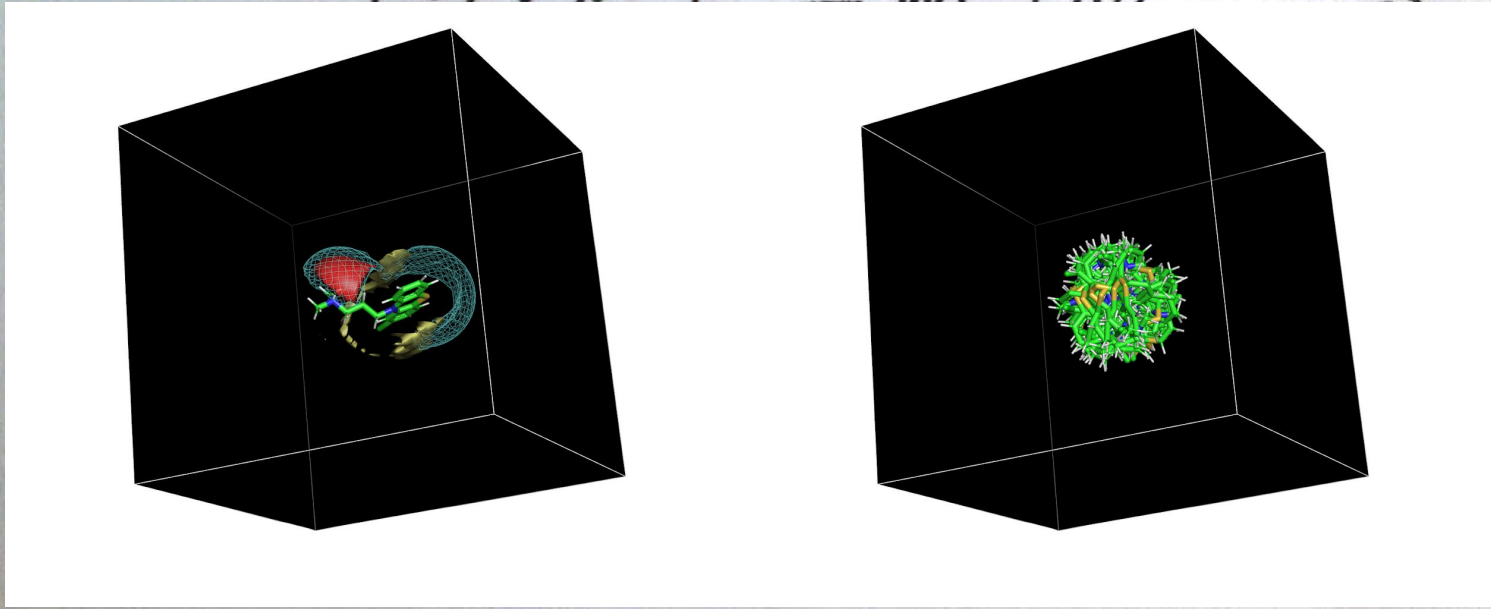


Note: in this case the training and validation sets were mixed so that different viewpoints of the same molecule were in training/validation, to allow the model to learn from the viewpoints

DeepGRID is alignment independent

Each molecule conformation centred within a grid cage 0,0,0 to 30,30,30

27 ‘Viewpoints’ generated by rotating the molecule around each axis



t's try it anyway...

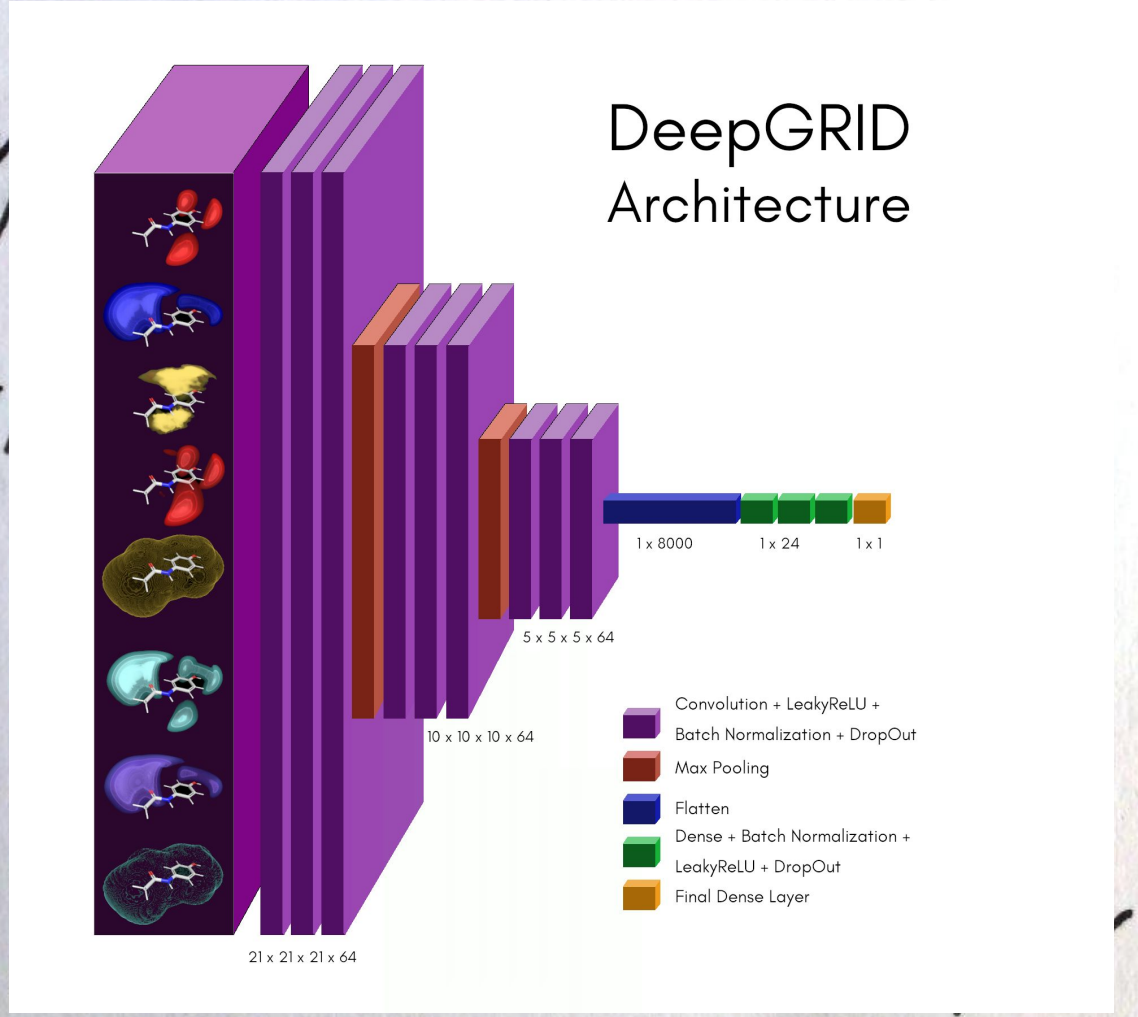
$$(i\gamma^\mu \partial_\mu - m)\psi = 0$$

MODEL

motion, want this, but $i\gamma^\mu$ not four

DeepGRID Model

- 3 convolutional layers, drop out and max pooling
  - extracting features and reducing the dimensionality
- Flattening layer
- 3 dense layers and drop out before the final dense layer



OTHER MODELS AND FEATURES

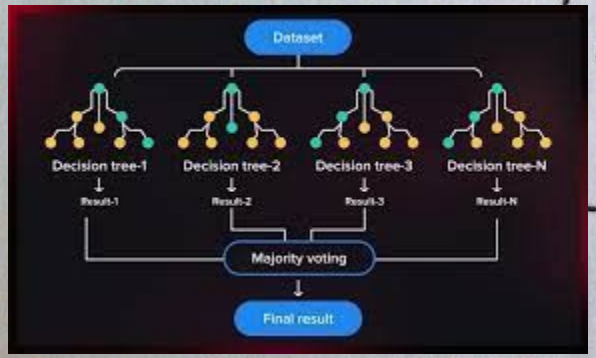
DeepGRID Hyperparameters optimization

Volsurf Descriptors

Descriptors	Probes*			Description
	OH2	DRY	O	
V	X			Molecular volume
S	X			Molecular surface
POL				Polarizability
MW				Molar mass
HB1–HB8			X	Hydrogen bonding
A				Amphiphilic moment
BV	X		X	Best volumes
W1–W8	X			Hydrophilic regions
ID1–ID8		X		Hydrophobic integy moment
Cw1–Cw8	X			Capacity factor
D1–D8		X		Hydrophobic regions
CP				Critical packing
LOG P				logarithm of partition coefficient
DIFF				Diffusivity

* Blank, other ways of calculation. For deatails, see reference Cruciani et al. (2000).

Random Forest Approach



- Each molecule conformation was used to calculate the VolSurf descriptors
- The VS model descriptors were removed (eg. LgBB and Caco2)
- A grid search was performed to optimize the hyperparameters and identify the best model scored using the validation set

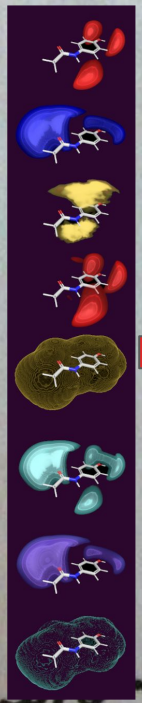
Partial Least Squares Approach

$$y_{nj} = \sum_{i=0}^{k} \beta_i x_{ni} + \varepsilon_{nj}$$

It is a linear relation but instead of the pure X variables we are using LV (Latent Variables) similarly to PCR (Principal Components Regression) but LV are build to “better correlate” also Y variable respect to PC (Principal Components).

- Each molecule conformation was used to calculate the VolSurf descriptors
- The VS model descriptors were removed (eg. LgBB and Caco2)
- A PLS model was generated and the number of components has been obtained looking for the best RMSE in the validation set while increasing the number of LV (Latente Variables)

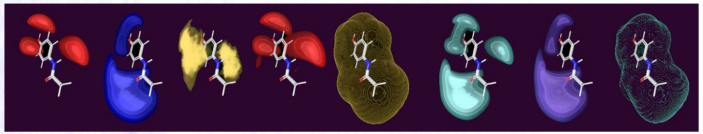
DeepGRID vs RF and PLS models



Volsurf3 Descriptors

Descriptors	Probes*			Description
	OH2	DRY	O	
V	X			Molecular volume
S	X			Molecular surface
POL				Polarizability
MW				Molar mass
HB1–HB8			X	Hydrogen bonding
A				Amphiphilic moment
BV	X		X	Best volumes
W1–W8	X			Hydrophilic regions
ID1–ID8		X		Hydrophobic integy moment
Cw1–Cw8	X			Capacity factor
D1–D8		X		Hydrophobic regions
CP				Critical packing
LOG P				logarithm of partition coefficient
DIFF				Diffusivity

* Blank, other ways of calculation. For deatails, see reference Cruciani et al. (2000).



Extracted features used by the dense layers

t's try it anyway...

$$(i\gamma^\mu \partial_\mu - m)\psi = 0$$

RESULTS

motion, want this, but $i\gamma^\mu$ not four

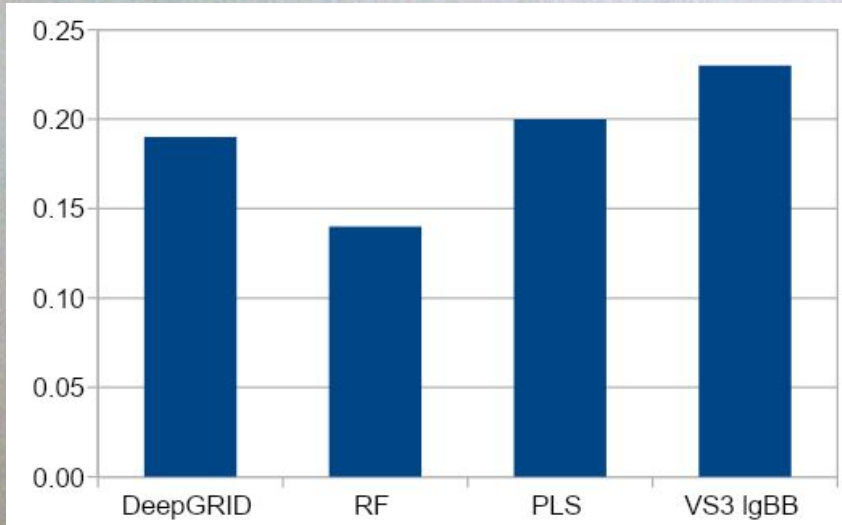
Removing CHEBI338620 as an outlier

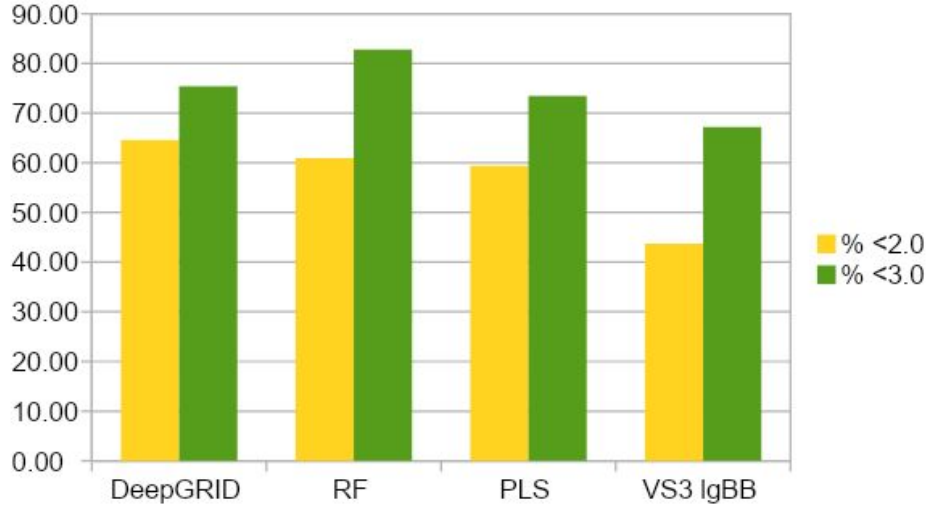
- CHEBI338620 has an reported experimental IgBB of -2.15
- However, it is very similar to Cimetidine which has shown limited BBB permeability
- There is also the possibility at extreme values that transporters are involved
- Without this, all models are better, but DeepGRID shows excellent performance

	MSE	GMFE	% <2.0	% <3.0
DeepGRID 75	0.24	3.87	63.6	74.2
RF	0.18	3.09	60.0	81.5
PLS	0.22	3.20	58.5	72.3
VS3 lgBB	0.27	3.77	43.1	66.2

Without CHEBI338620	MSE	GMFE	% <2.0	% <3.0
DeepGRID 75	0.19	2.79	64.6	75.4
RF	0.14	2.34	60.9	82.8
PLS	0.20	2.97	59.4	73.4
VS3 lgBB	0.23	3.18	43.8	67.2

Removing CHEBI338620 as an outlier





Lower is better

Higher is better

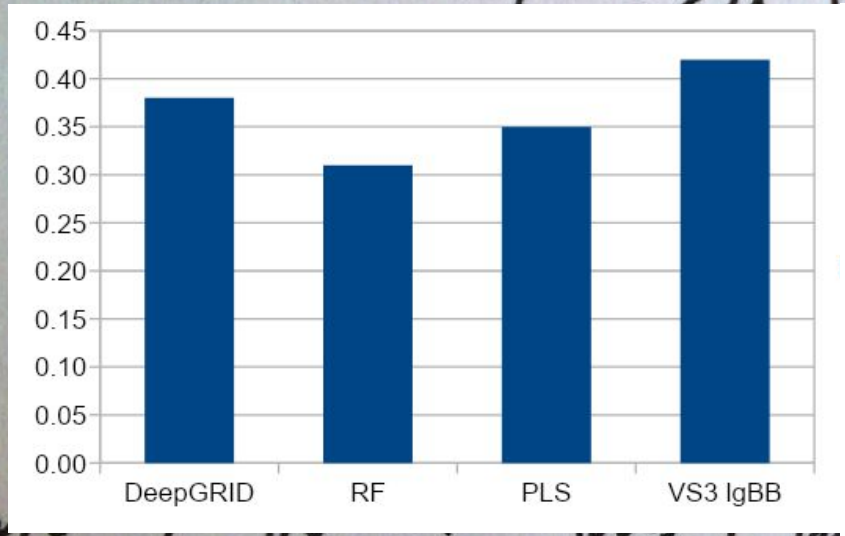
Light-IgBB-416 dataset is more diverse

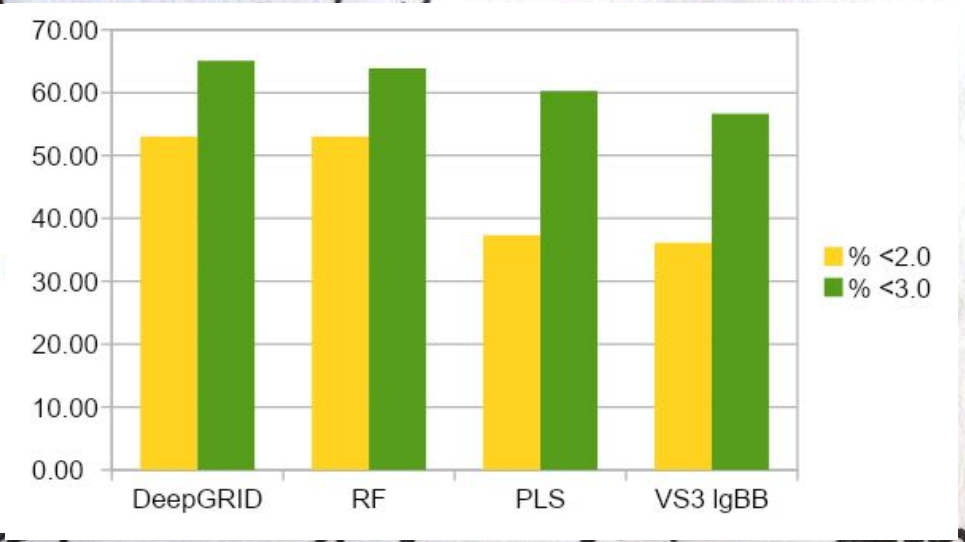
More diverse → more difficult→ all approaches give less accurate models

	MSE	GMFE	% <2.0	% <3.0
DeepGRID 75	0.38	5.04	53.0	65.1
RF	0.31	4.27	53.0	63.9
PLS	0.35	4.79	37.4	60.2
VS3 IgBB	0.42	7.78	36.1	56.6

Light-IgBB-416 dataset is more diverse

More diverse → more difficult→ all approaches give less accurate models



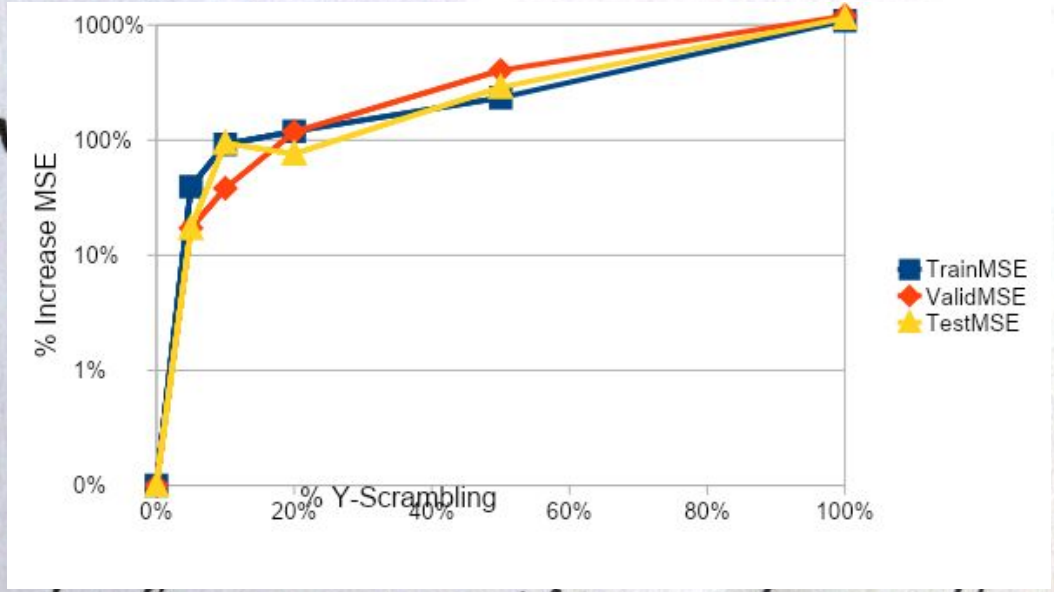


Lower is better

Higher is better

DeepGRID gives a robust model

- Y-Scrambling the data affects the model, ie. It is not overfitting
- At 5% scrambling the Test MSE is only 17% worse, hence the approach is relatively robust to erroneous data
  -

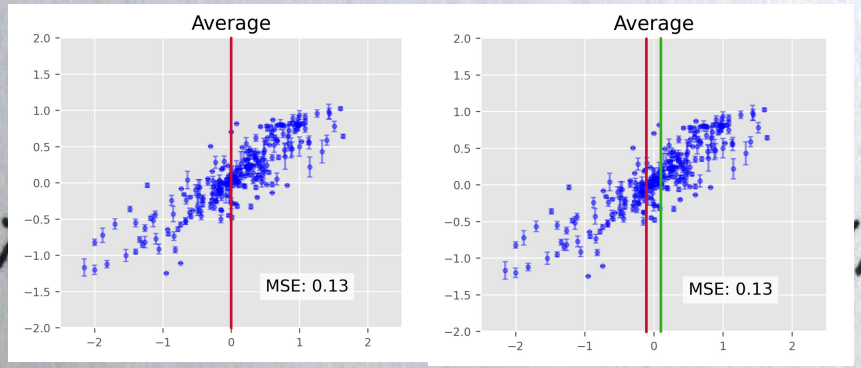


Initial Summary

- The DeepGRID model has successfully extracted relevant features from the raw GRID MIFs and given a good model when compared to standard approaches using the hand-crafted VolSurf descriptors
- Random Forest + VolSurf descriptors slightly better overall than all approaches

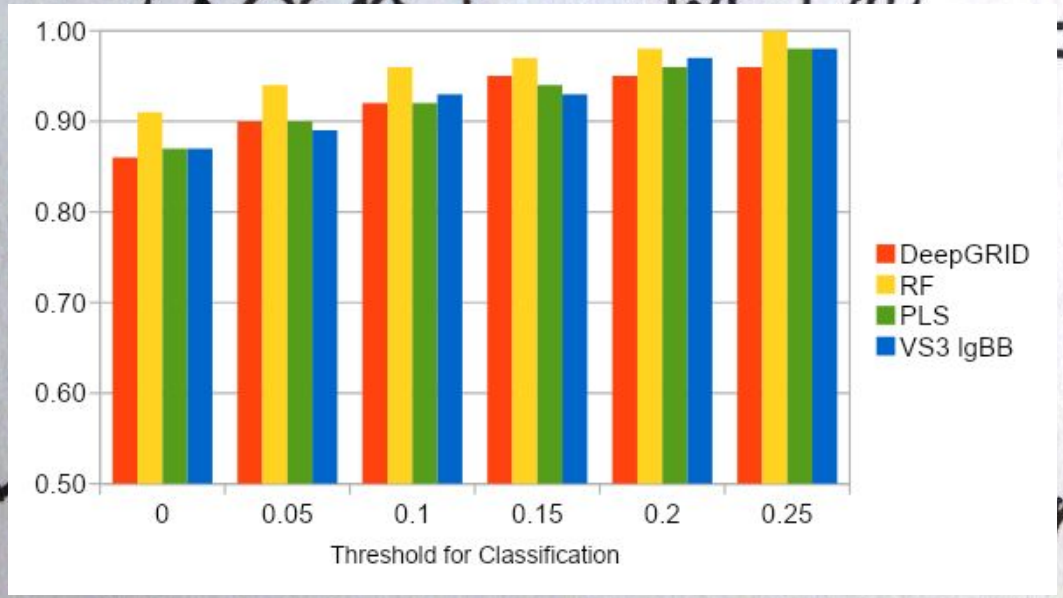
Regression → Classification

- The regression models for described can also be used for classification (BBB +/-)
- Compounds with experimental lgBB close to 0.0 may be ambiguous and misclassified
- In this case we measured the ROC AUC at varying thresholds on the **Test**



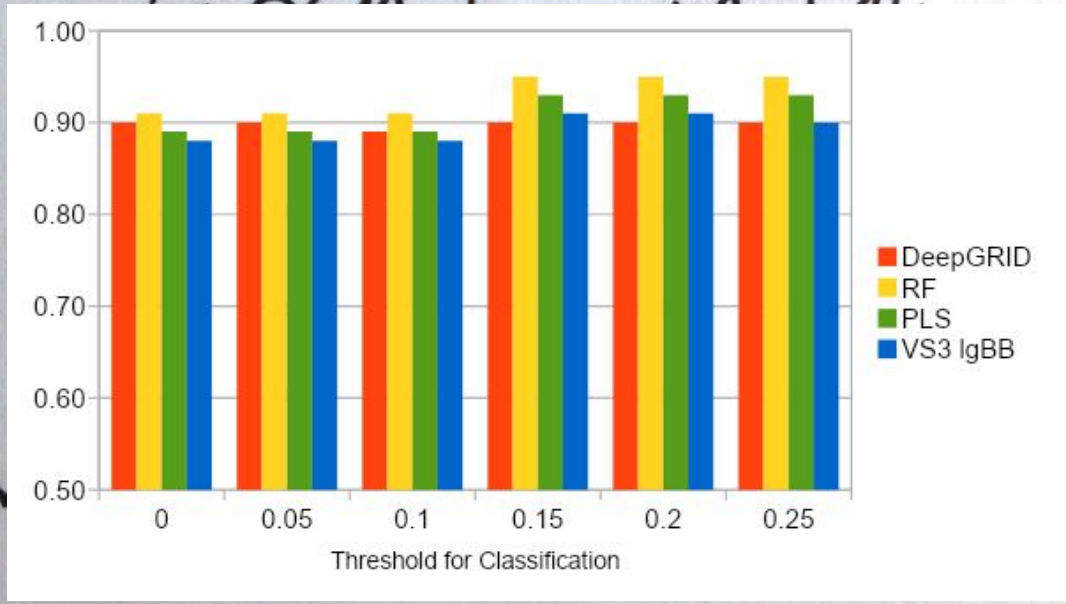
Classification: VS-IgBB-332 model

- At a minimal threshold of 0.1, all models predict with >90% accuracy
- The RF model is slightly better



Classification: Light-IgBB-416 model

- At minimal threshold of 0.1, all models predict with ~90% accuracy
- All models are fairly equal

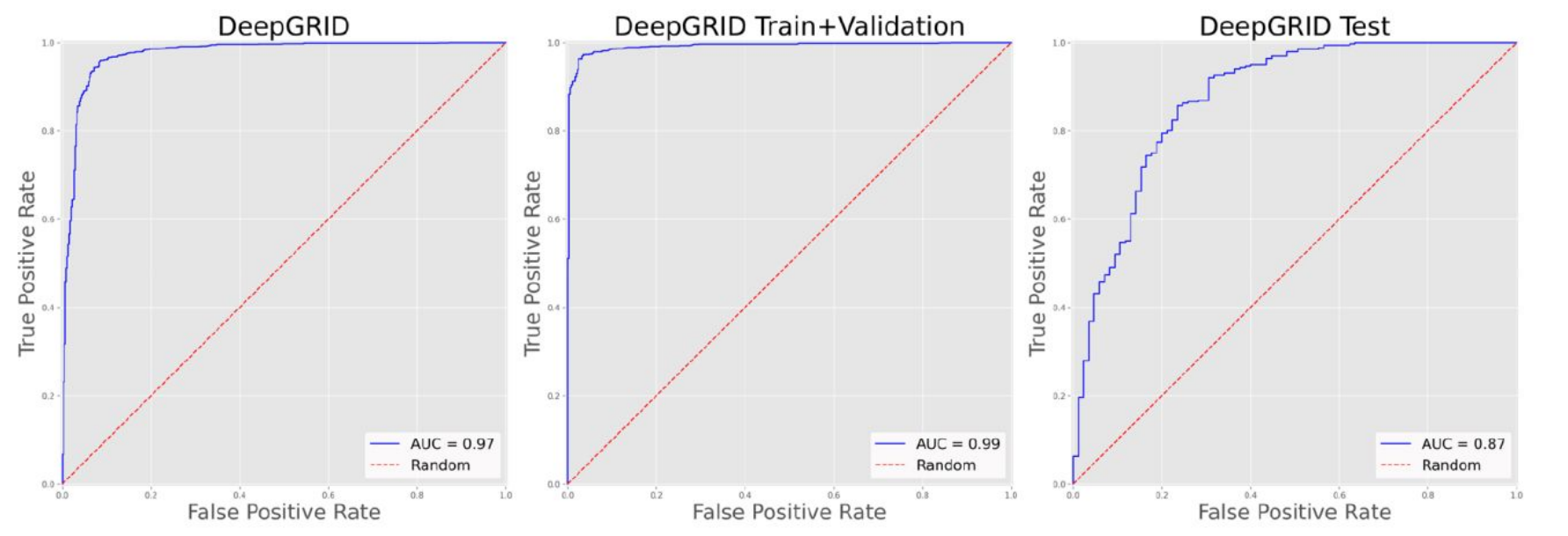


Classification Models - Light-IgBB-2105 dataset

- New classification models were built using DeepGRID and Random Forest (with hyperparameter optimization)
- Initial attempts with DeepGRID kept stalling during learning
- Potentially due to data imbalance?
- The BBB- cpds were artificially augmented to bring the balance to 0.5:1
  - successful learnin

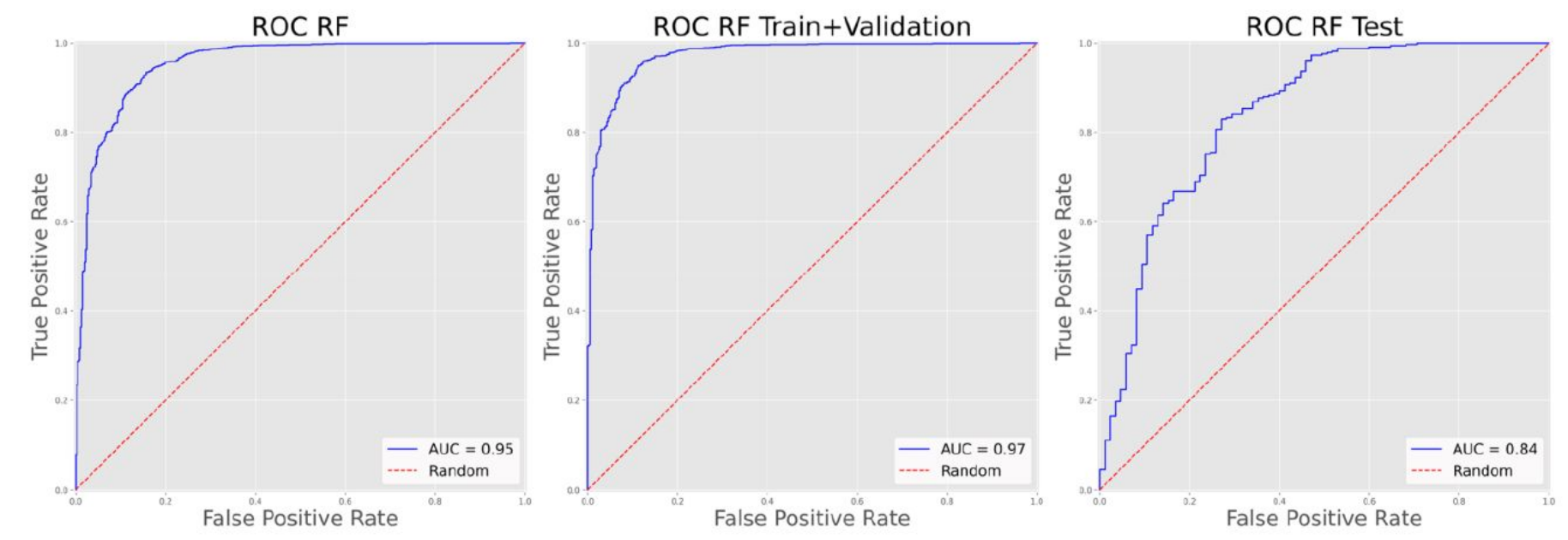
DeepGRID Classification Models - Light-IgBB-2105 dataset

AUC Full Set: **0.97** Test Set: **0.87**



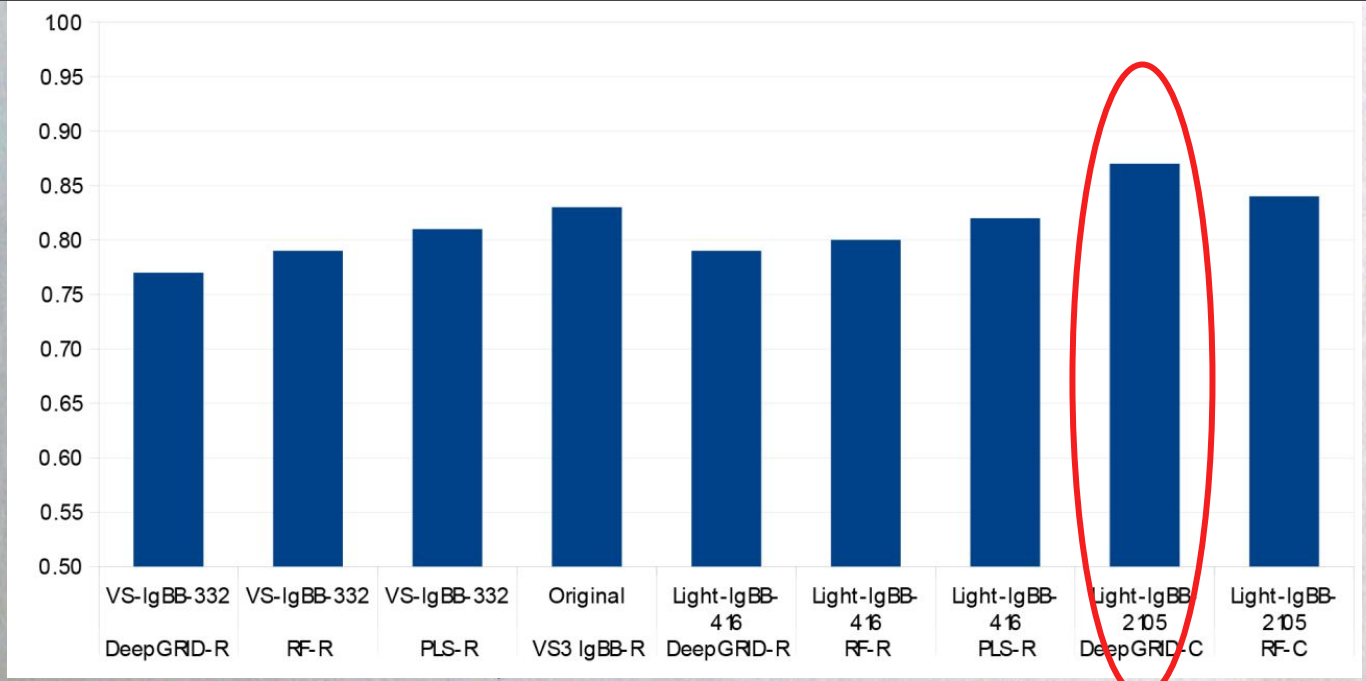
RF Classification Models - Light-IgBB-2105 dataset

AUC Full Set: **0.95** Test Set: **0.84**



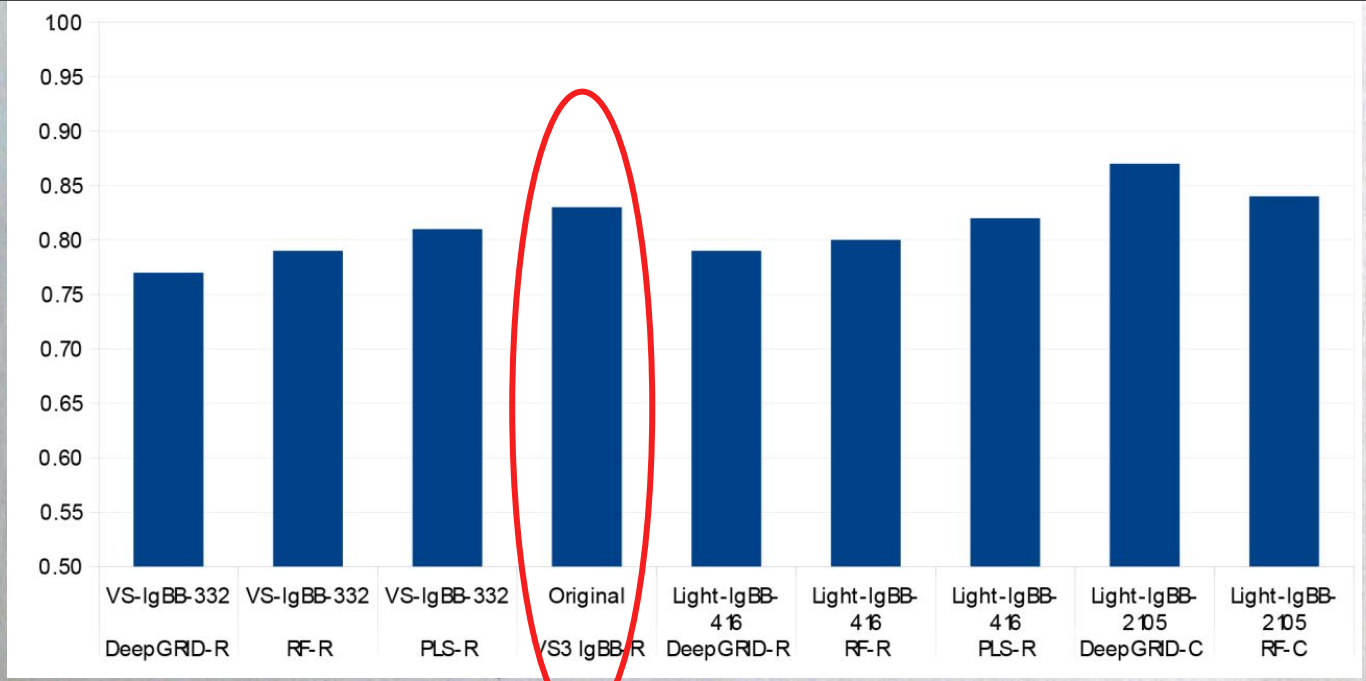
DeepGRID model the best for classification

All models classification performance (ROC-AUC) on the 2105 dataset



VolSurf lgBB PLS model does a good job

All models classification performance (ROC-AUC) on the 2105 dataset



- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

A Formula search

Methods, such as random forest (RF) or neural network (NN), are very efficient 36 but not always transparent, partially blurring the comprehension of the role played by the input variables in the final results

- Improvements toward the interpretability of such “black-box” ML models have been made through additional methodologies, such as model-agnostic methods (i.e., permutation feature importance)
- A ML-based approach to build sets of features (or descriptors) starting from a given set of basic variables (e.g., atomic properties), subsequently used to construct LR models (or formulas)

Inspired by the original work of Ghiringhelli et al. prediction of the difference in energy between RS and ZB; from that optimization, a classification of the most stable crystal structure between RS (rocksalt) and ZB (zinc blende) for semiconductor AB binary compounds naturally derives **(full dataset is made of 82 compounds)**

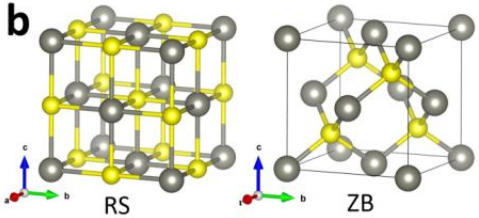
Udaykumar Gajera, Loriano Storchi, Danila Amoroso, Francesco Delodovici, Silvia Picozzi "Towards machine learning for microscopic mechanisms:a formula search for crystal structure stability based on atomic properties" Journal of Applied Physics, DOI: 10.1063/5.0088177 (2022)

A Formula search

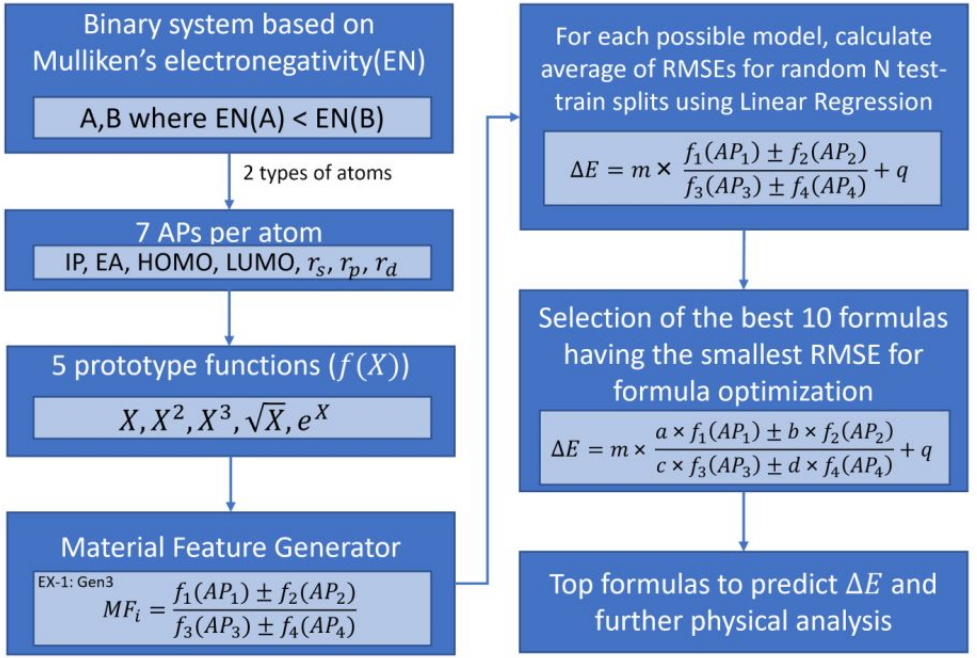
a

7 Atomic Properties (APs)	
IP	Ionization potential
EA	Electron Affinity
HOMO	Highest occupied level
LUMO	Lowest unoccupied level
r_s	radii of s orbital
r_p	radii of p orbital
r_d	radii of d orbital

b



c



(a) Basic atomic properties (APs) used to construct the material features. (b) Crystal structures of RS and ZB (plot made using the VESTA tool). 62 Gray (yellow) spheres represent A (B) atoms. (c) Workflow for formula construction, machine-learning methodology, validation, and MF selection.

A Formula search

GEN1: combine two prototype functions in the numerator, forcing them to belong to the same kind of APs, which is both “spatial”-like or both “energy”-like; one prototype function is at the denominator with the only constraint to be non-zero

GEN2: combine two prototype functions with the same kind of APs at the numerator and a single prototype function at the denominator with an argument of a different kind with respect o the numerator ones. For instance, if AP_1 in $f_1(AP_1)$ and AP_2 in $f_2(AP_2)$ are “energy” terms (i.e., EA or HOMO), then AP_3 must be a “spatial” term (i.e., r_n)

$$MF = \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3)}$$

A Formula search

GEN3: combine two prototype functions at both the numerator and denominator without any constraints,

$$MF = \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3) \pm f_4(AP_4)}$$

GEN4: combine two prototype functions with the same physical dimensions at both the numerator and denominator

$$MF = \frac{f_1(AP_1) \bigstar f_2(AP_2)}{f_3(AP_3) \bigstar f_4(AP_4)},$$

$\bigstar = + - \times \div$

A Formula search

Formula	avg (RMSE)	RMSE	R^2	Success rate (%)	Generator type
$0.117 \times \frac{EA(B)-IP(B)}{r_p(A)^2} - 0.342$	0.1455	0.1423	0.89	89	1D descriptor[55]
$-0.751 \times \frac{r_p(B)^3-\exp[r_s(B)]}{r_p(A)^2} - 0.317$	0.1296	0.1193	0.92	90	*GEN*1
$0.285 \times \frac{\sqrt{\lvert IP(B)\rvert}+\sqrt{\lvert EA(A)\rvert}}{r_p(A)^2} - 0.387$	0.1367	0.1309	0.91	91	*GEN*2
$0.774 \times \frac{r_p(B)+\sqrt{\lvert r_d(A)\rvert}}{r_p(A)^3+r_p(B)^3} - 0.303$	0.0995	0.0963	0.95	94	*GEN*3
$1.155 \times \frac{r_s(B)+r_s(A)}{r_p(B)^3+r_p(A)^3} - 0.368$	0.1103	0.1058	0.94	96	*GEN*4

1D formulas, along with related statistics: avg(RMSE) denotes the root mean squared error for average over 1000 random train-test splits of dataset. Instead, the RMSE is the root mean squared error for the entire dataset as training and test. Similarly, the R^2 values are calculated considering the entire dataset, and they show the quality of fit between predicted and actual values. The success rate (in percent) shows how many RS or ZB phases out of 82 have been correctly identified by the descriptor. The "Generator type" column indicates the different generators used to produce the corresponding formula. RMSEs are in eV.

A Formula search

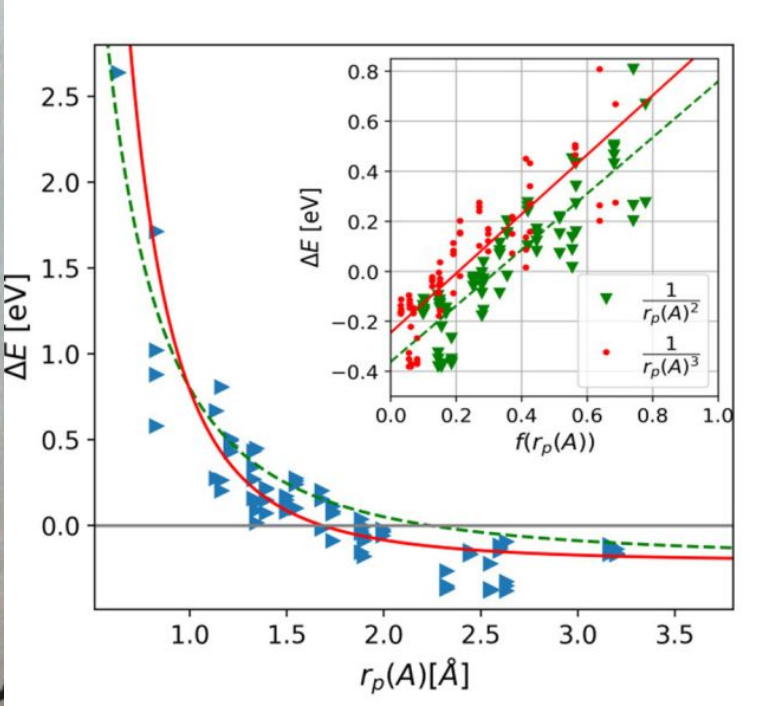
$$\Delta E = m \times \frac{a \times f_1(AP_1) \bigstar b \times f_2(AP_2)}{c \times f_3(AP_3) \bigstar d \times f_4(AP_4)} + q,$$

GRID search, for each set of weight coefficients generated during the grid search, we also run the linear regression. Thus, we are performing a proper formula optimization, as at each step of the grid search, we are updating both the weight coefficients as well as the slope and intercept coming from the LR

Formula	avg (RMSE)	RMSE	R^2	Success rate (%)	Generator type
$0.127 \times \frac{0.800 \times EA(B) - 1.000 \times IP(B)}{1.110 \times r_p(A)^2} - 0.352$	0.1457	0.1419	0.89	89	1D descriptor[55]
$-1.870 \times \frac{0.801 \times \sqrt{r_p(B)} - 0.606 \times \exp[r_p(A)]}{1.010 \times r_p(A)^3} - 0.968$	0.1191	0.1143	0.93	91	*GEN*1
$0.477 \times \frac{0.876 \times \sqrt{\lvert HOMO(B)\rvert} + 0.468 \times \sqrt{\lvert LUMO(B)\rvert}}{1.110 \times r_p(A)^2} - 0.372$	0.1340	0.1296	0.91	91	*GEN*2
$1.609 \times \frac{0.642 \times r_p(B) + 0.502 \times \sqrt{\lvert r_d(A)\rvert}}{1.170 \times r_p(A)^3 + 1.170 \times r_p(B)^3} - 0.309$	0.0991	0.0961	0.95	94	*GEN*3
$1.207 \times \frac{0.878 \times r_s(B) + 0.200 \times r_p(A)}{0.512 \times r_p(B)^3 + 0.610 \times r_p(A)^3} - 0.359$	0.1045	0.1016	0.94	99	*GEN*4

1D formulas after the optimization step, along with related statistics. Notation as in Table I. RMSEs are in eV.

A Formula search



The final outcome of our procedure is a transparent formula, not necessarily of easy mathematical formulation, but revealing which part of the input mostly affects the output, i.e., allowing the identification of the main driving physical feature

Interestingly, our results reveal the size of the A cation to play a leading role in the phase stabilization; in fact, the r_n(A) radius appears in the best-performing formulas more frequently than the other basic atomic properties

Data fit functions are also shown, using proportionality to rp(A)$^{-2}$ and rp(A)$^{-3}$ via a green dashed line and a red straight line, respectively.

A Formula search

Generator	Total Number of generated formulas	Elapsed time (s) for 1D formula construction	Elapsed time (s) for formula optimization
GEN1	106400	5117.32	180.84
GEN2	67840	3338.93	181.54
GEN3	1091200	51821.54	420.52
GEN4	278106	13237.39	418.62

Time needed to generate the best 1D formula and perform its optimization. All the calculations have been performed in a PC equipped with an Intel Core i5-8500 processor and 16 GiB of RAM.

- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

Random Forrest and Permutation Feature Importance

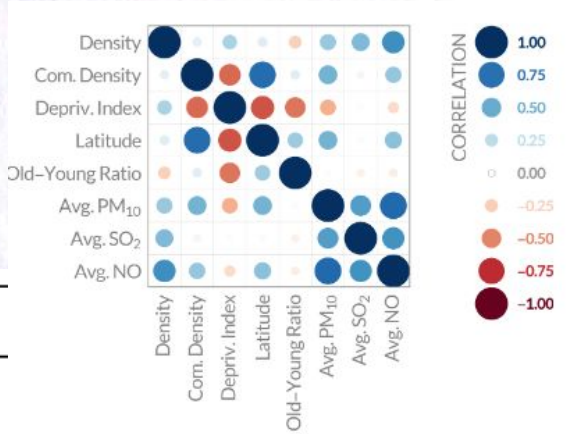
Use the RF model not for prediction purpose but to detect how much a feature is important respect to the others. Two ingredients:

- The permutation feature importance is defined to be **the decrease in a model score when a single feature value is randomly shuffled**. This procedure breaks the relationship between the feature and the target, thus the drop in the model score is indicative of how much the model depends on the feature
- Random forests or random decision forests is **an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time**

Leonardo Aragao, Elisabetta Ronchieri, Giuseppe Ambrosio5, Diego Ciangottini, Sara Cutini, Cristina Duma, Pasquale Lubrano, Barbara Martelli, Davide Salomoni, Giusy Sergi, Daniele Spiga, Fabrizio Stracci, Loriano Storchi "Air quality changes during the COVID-19 pandemic guided by robust virus-spreading data in Italy",to Air Quality, Atmosphere & Health, DOI: 10.1007/s11869-023-01495-x (2024)

Features

Feature name	Description
Population Density	Population divided by province's area.
Commuting Density	Percentage of commuters over population [8].
Deprivation Index	Represents the multidimensionality of the social and material deprivation concept [29] (calculated for the year 2012).
Latitude	North–south geographic coordinate regarding the province's capital.
Old-Young Ratio	Number of individuals aged 20 or less over the ones aged 65 and over.
Avg. PM_{10}	Average concentration of PM_{10} during the whole study period.
Avg. NO	Average concentration of NO during the whole study period.
Avg. SO_2	Average concentration of SO_2 during the whole study period.

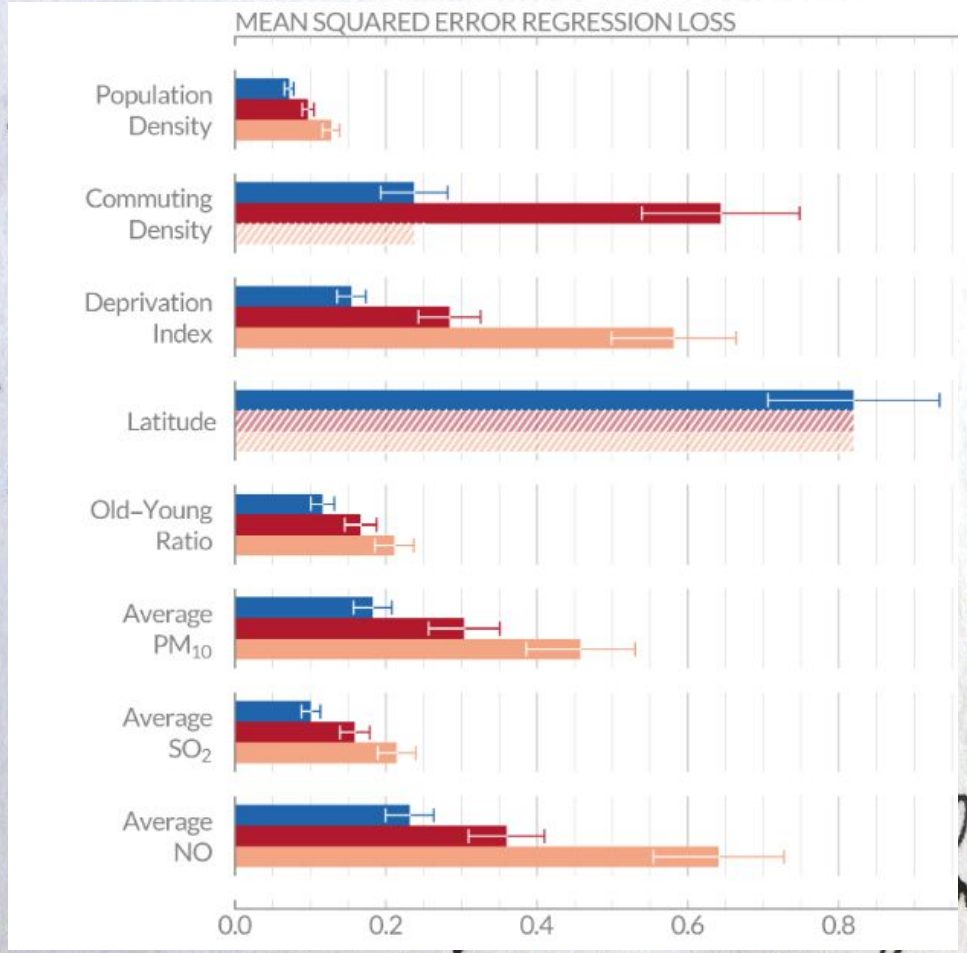


Results

104 Italian provinces analysed applying the Permutation Feature Importance Analysis to a set of different Random Forest models

The role of the pollutants seems not the most important

Details	RMSE	R^2
All features	0.320	0.950
Latitude Removed	0.341	0.943
Latitude and Comm. Density removed	0.362	0.936



- ML Introduction
  - Deep Learning and Neural Network: Introduction
- DeepGRID Test Case: Blood Brain Barrier Permeation
  - GRID MIFs: Introduction
  - DeepGRID
- Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties
- Possible correlation between pollutants and COVID-19 cases
- Conclusions

Conclusions

- Deep Learning successfully used with GRID MIFs → DeepGRID
  - Regression models among the best
  - Larger Classification model the best (Test Set AUC: 0.87, Overall AUC: 0.97
- Formula generator
  - Approach can be used also with small datasets
  - New generators can be easily plugged including different constraints
  - The final results is a mathematical formula human readable
- Next step: Generative AI