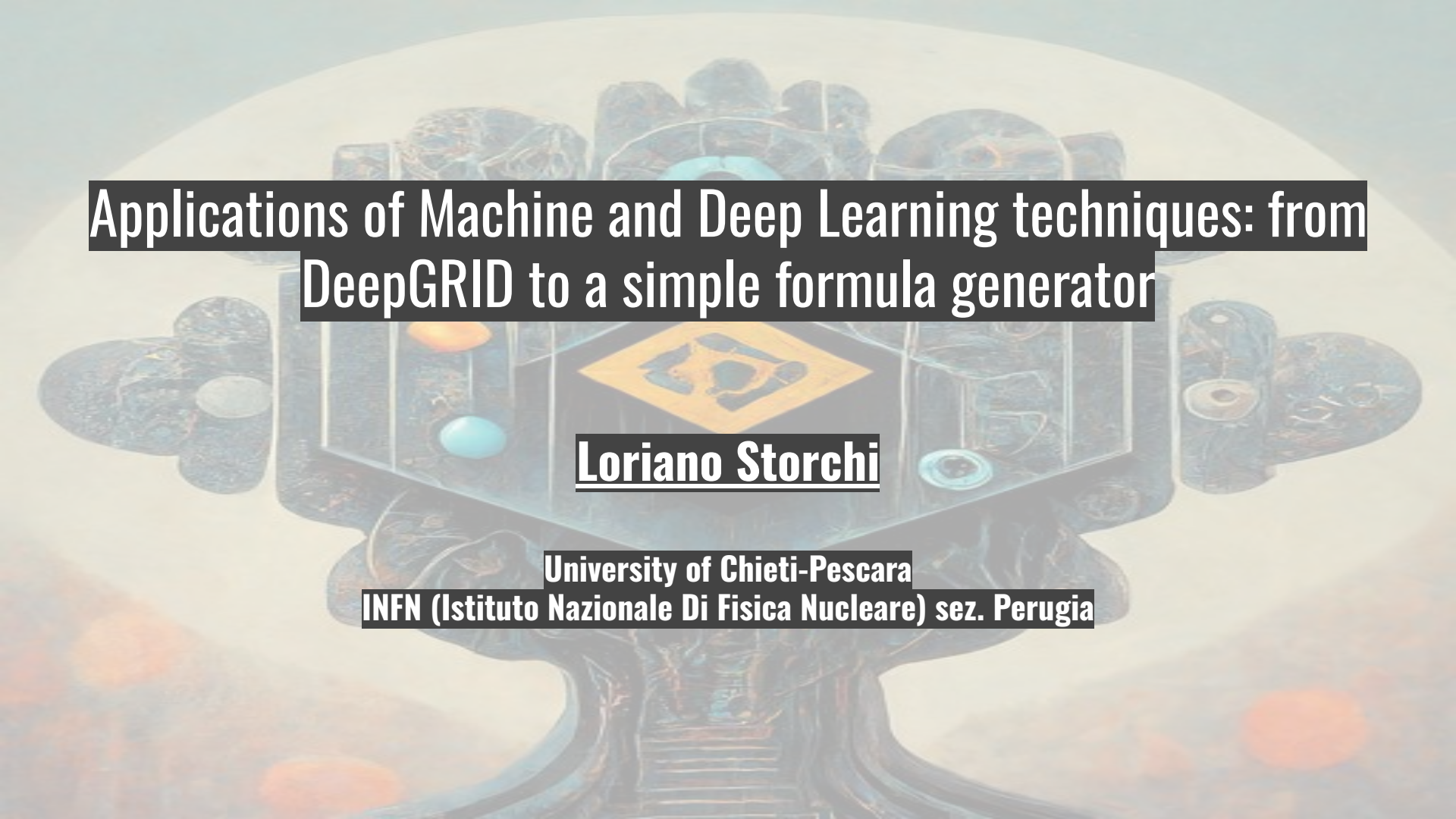


Applications of Machine and Deep Learning techniques: from DeepGRID to a simple formula generator

Loriano Storchi

University of Chieti-Pescara
INFN (Istituto Nazionale Di Fisica Nucleare) sez. Perugia

My activities

- Four Component Dirac-Kohn-Sham Theory (BERTHA code)
  - CNR and UNIPG
- Machine Learning and Chemoinformatics
  - UNIPG and MolDiscovery
- HEP (High Energy Physics) - ML techniques and FPGA (Field-programmable gate array) and Cloud Computing
  - INFN and CERN
- Bio and Chemoinformatics
  - UNICH

- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

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

Machine learning techniques can be divided into two foremost types:

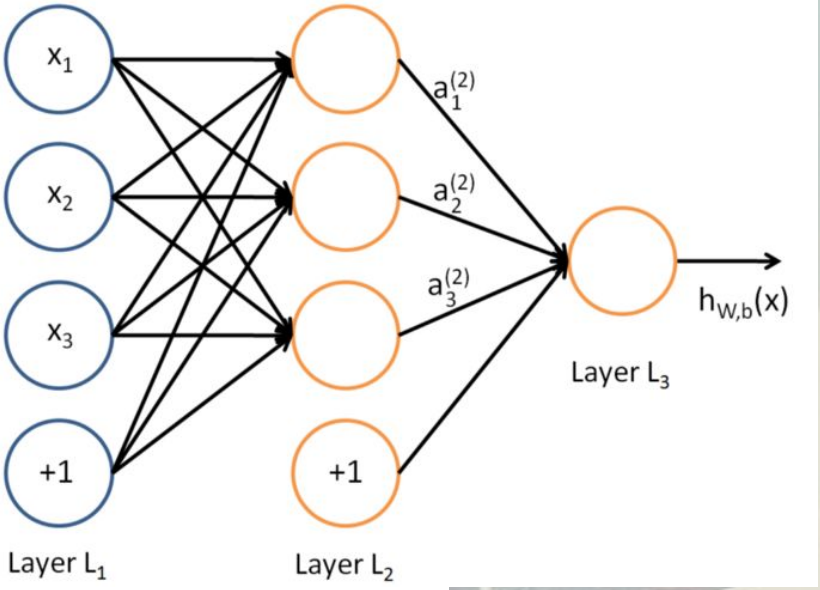
- **Unsupervised**: find hidden patterns or intrinsic structures in data. They are used to draw inferences from data sets consisting of input data without labeled responses (i.e. clustering algorithms)
- **Supervised**: used when you want to predict or explain the data you possess. A supervised algorithm 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)

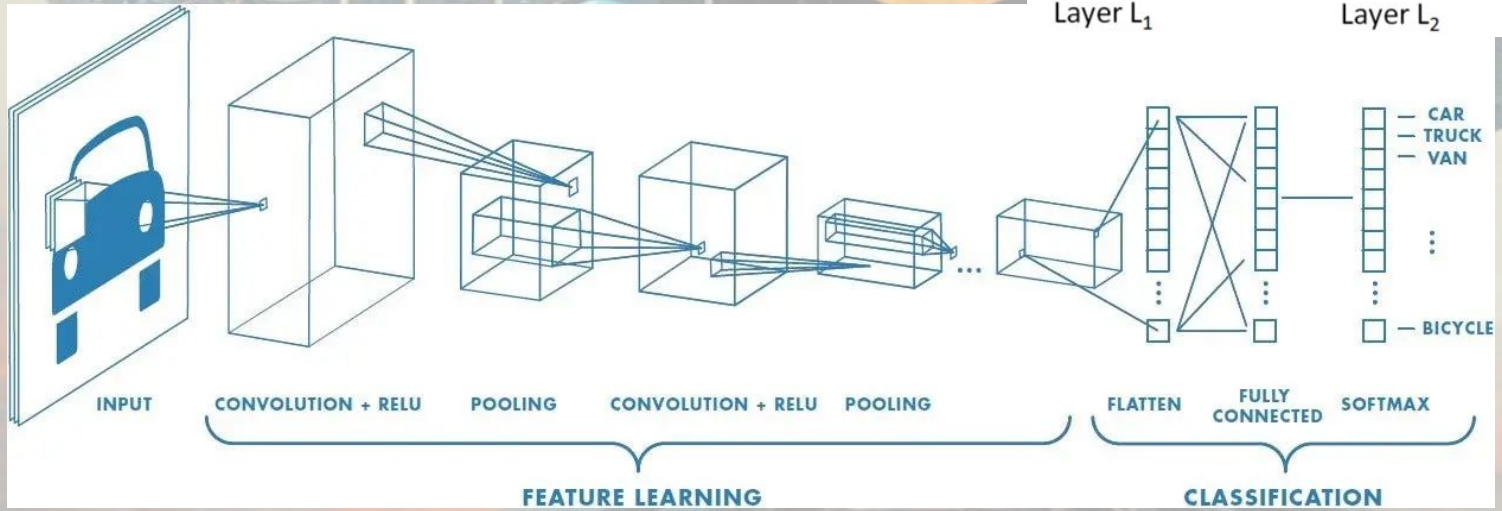
Tommaso Tedeschi, Marco Baioletti, Diego Ciangottini, Valentina Poggioni, Daniele Spiga, Loriano Storchi, Mirco Tracolli, "Smart Caching in a Data Lake for High Energy Physics Analysis", Journal of Grid Computing, DOI: 10.1007/s10723-023-09664-z (2023)

- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

Neural Network abd CNN

Deep learning is a subset of machine learning that uses artificial neural networks with multiple layers to extract higher-level features from raw input data.





- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients

- Rate coefficients for vibrational energy transfer are calculated for collisions between molecular nitrogen and hydrogen in a wide range of temperature and of initial vibrational states
  - The calculations were performed by a mixed quantum-classical method

ML Goal Predict rate coefficients for vibrational energy transfer processes involving specific initial vibrational states, which are computationally expensive to calculate directly.

Qizhen Hong, Loriano Storchi, Massimiliano Bartolomei, Fernando Pirani, Quanhua Sun, Cecilia Coletti, "Inelastic N2+H2 collisions and quantum-classical rate coefficients: large datasets and machine learning predictions" The European Physical Journal D, DOI: 10.1140/epjd/s10053-023-00688-4 (2023

N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients

We used an tested two possible approaches:

- Neuronal Network (NN)
- Gaussian Process Regression (GPR):
  - Non-parametric, Bayesian approach to regression.
  - Flexible models that work well on small datasets
  - Start from the assumption that f(x) and f(y) are normally distributed with some meand and some covariance being x known points (training) and y unknown points (test)
  - Make predictions based on the similarity between data points.
  - Kernel function defines the similarity measure between points.
  - Hyperparameters of the kernel function are learned from the data.

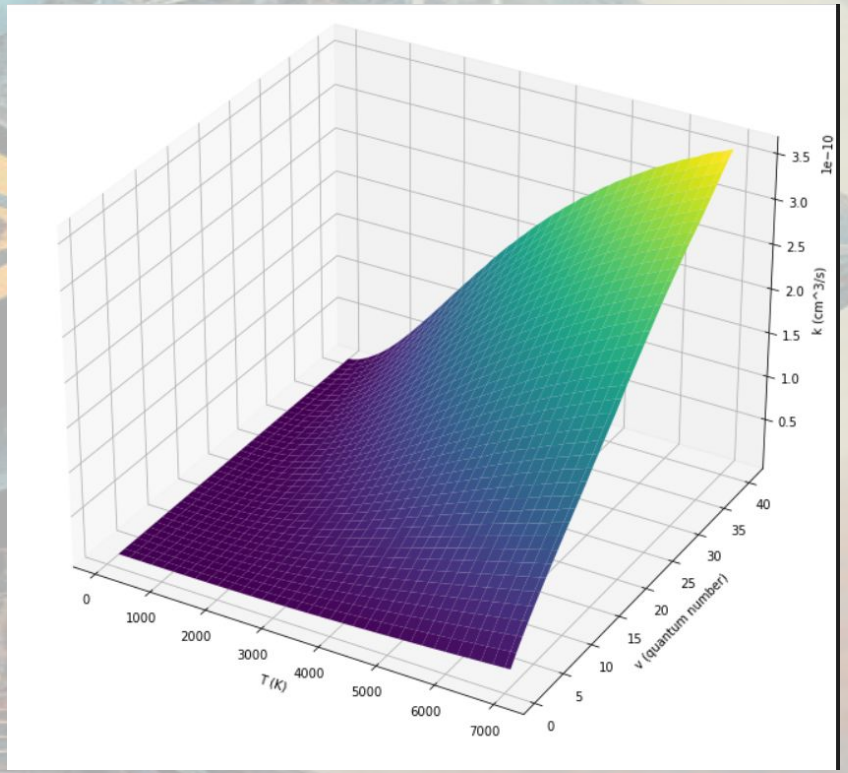
N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients

$$N_2(v) + H_2(0) \longrightarrow N_2(v - \Delta v) + H_2(0), \quad \Delta v = 1, 2, 3.$$

$\log_{10}(k)$ is the label

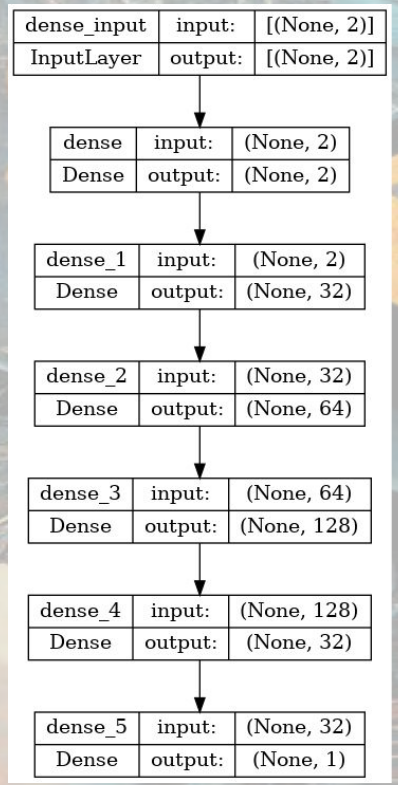
v, T are the two features

We want to test the performances of ttwo models NN and GPR



N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients

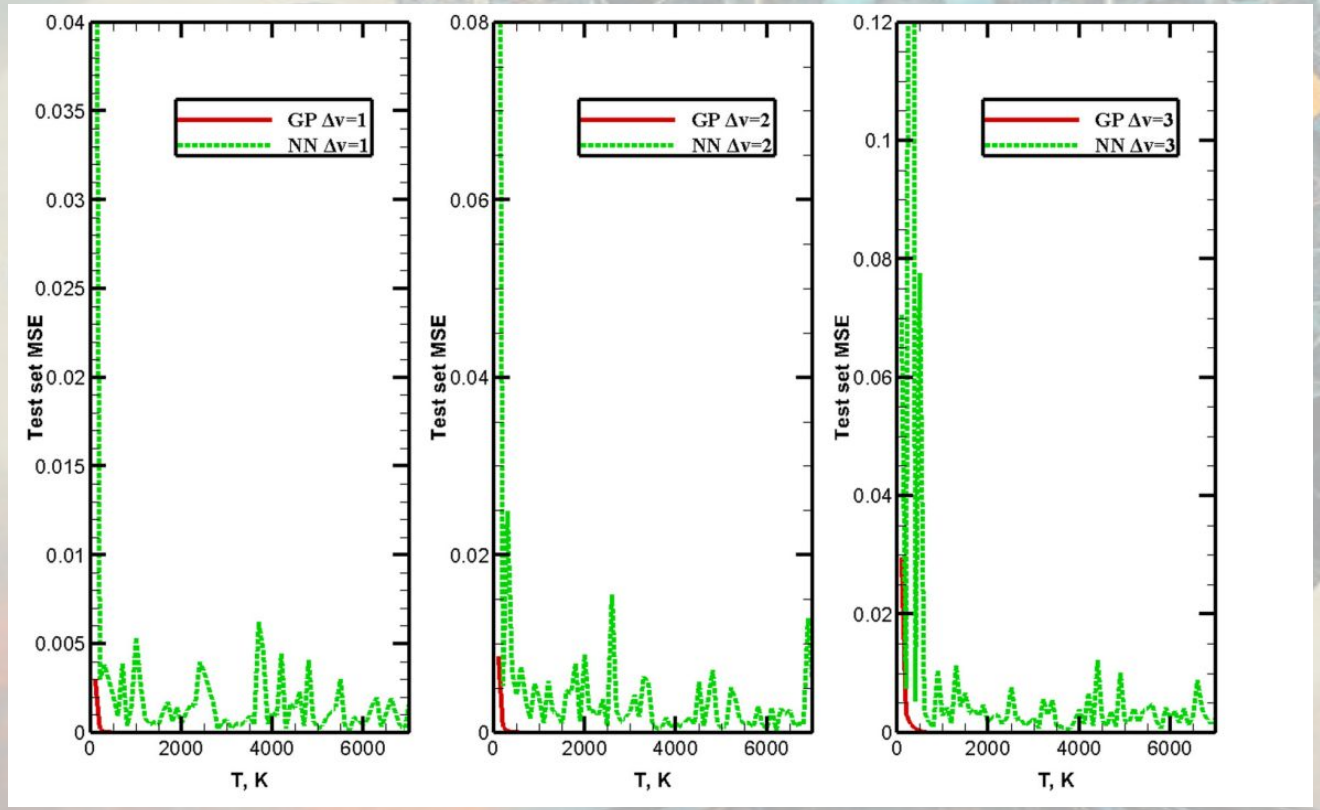
NN model unsinf Linear activation in input and output and ReLU



GPR using Matern Kernel
v = 5/2

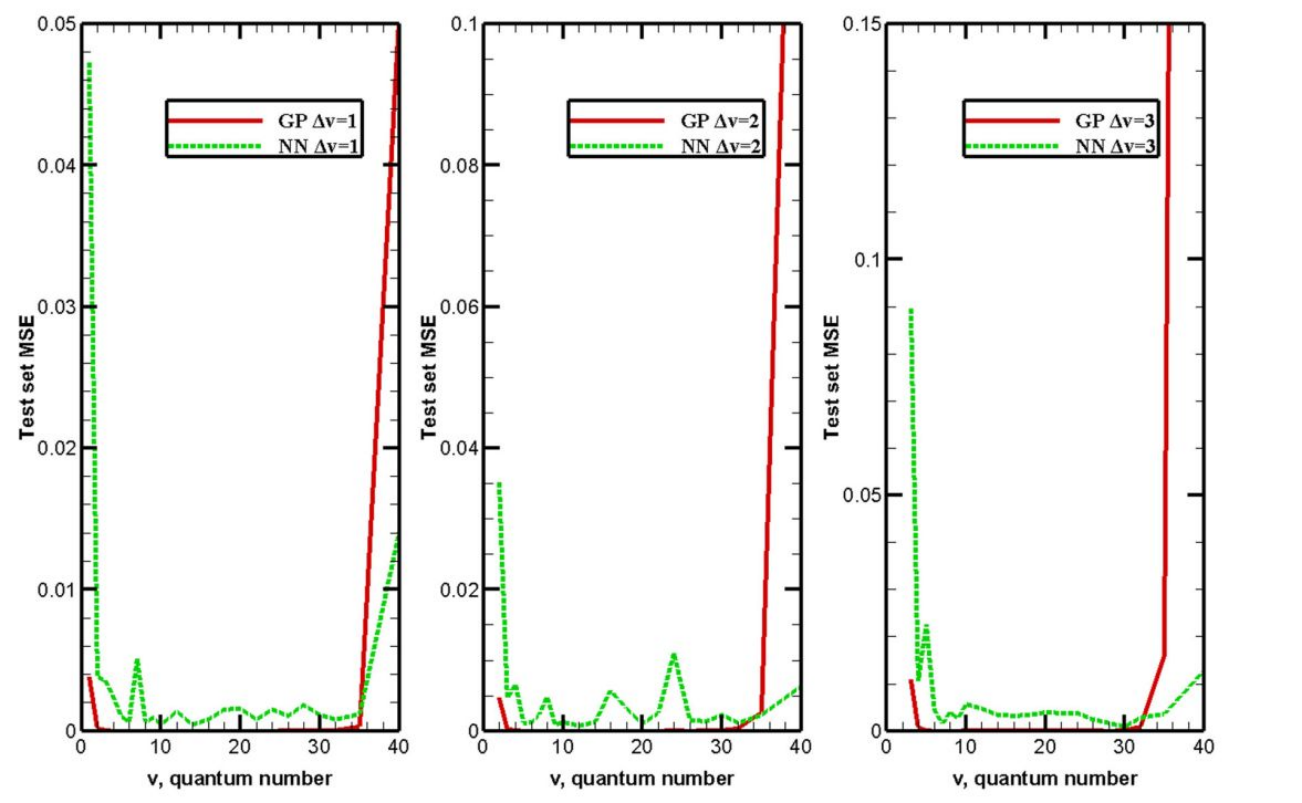
$$k(x_i, x_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}}{l}d(x_i, x_j)\right)$$

N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients



Test set MSE values as a function of temperature: $\log_{10}$ (k) values corresponding to a specific temperature T were removed from the training set and constitute the test set. The three panels correspond to processes (5) with Δv = 1, 2, 3, respectively

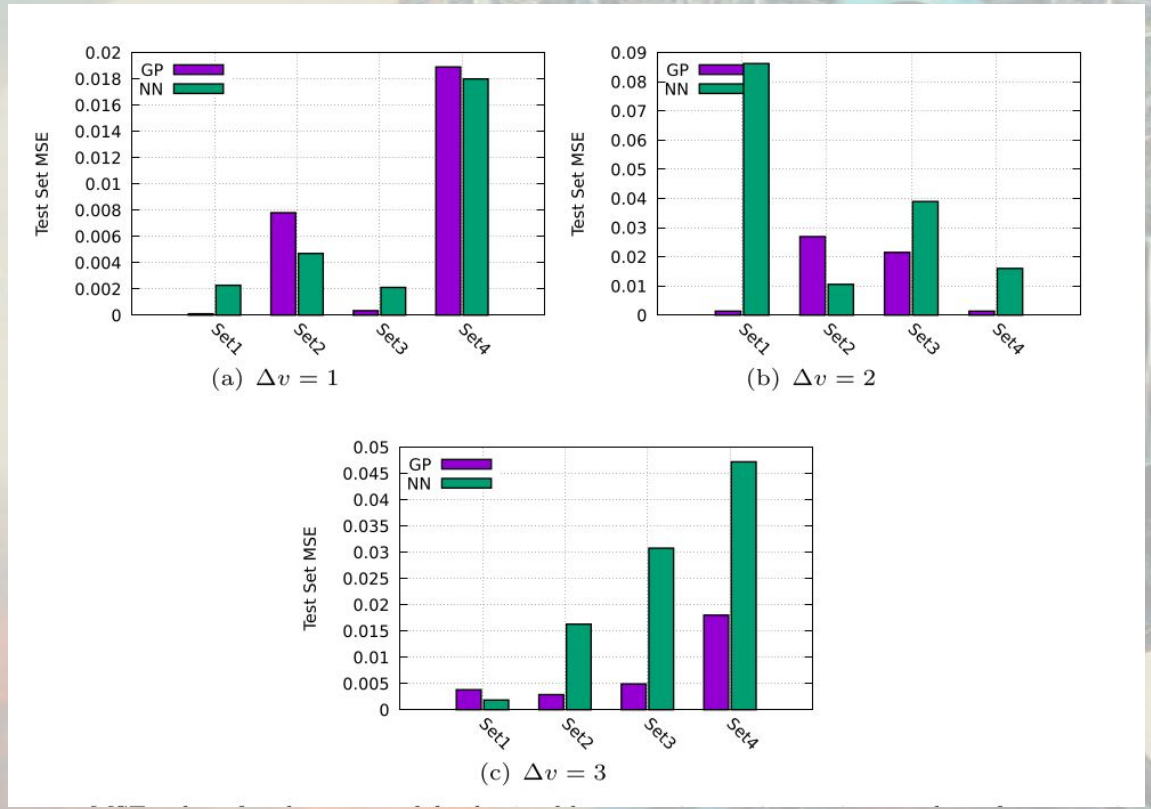
N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients



E values as a
nitial
quantum
g_{10} (k) values
ng to a
ere removed
ining set and
he test set.
anels
to processes

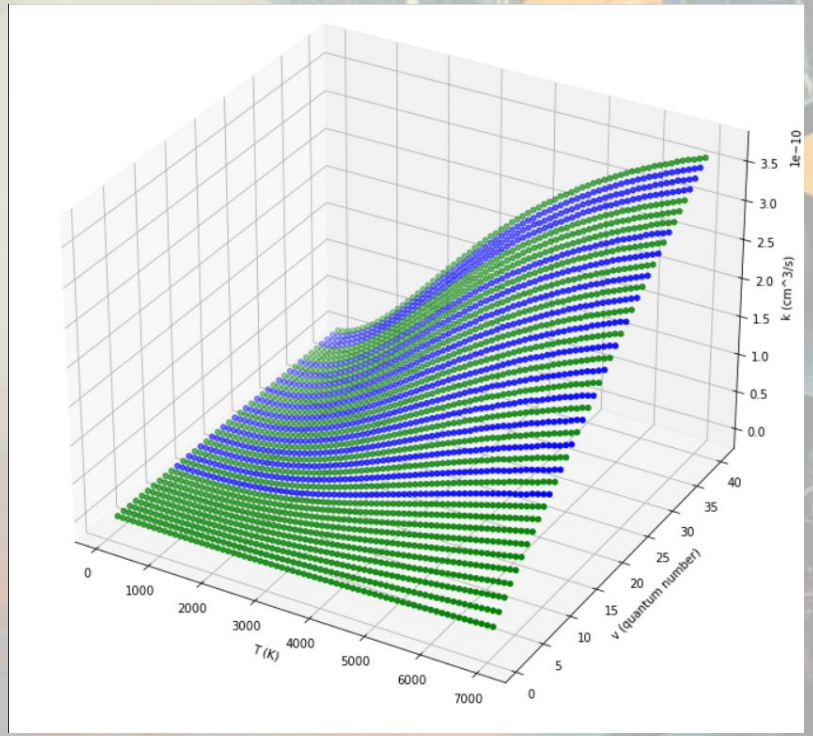
respectively

N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients



The test set MSE values for the two models obtained by removing an increasing number of systematically selected points, corresponding to specific v values, from the training set, i.e., Set1, removed v = [2; 4; 6; 8; 10; 14; 18; 22; 26; 30; 35], Set2, removed v = [1; 3; 5; 7; 9; 12; 16; 20; 24; 28; 32; 40], Set3, removed v = [2; 3; 5; 6; 8; 9; 12; 14; 18; 20; 24; 26; 30; 32], Set4, removed v = [1; 2; 4; 5; 7; 8; 10; 12; 16; 18; 22; 24; 28; 30; 35; 40]. The three panels correspond to processes (5) with Δv = 1, 2, 3, respectively

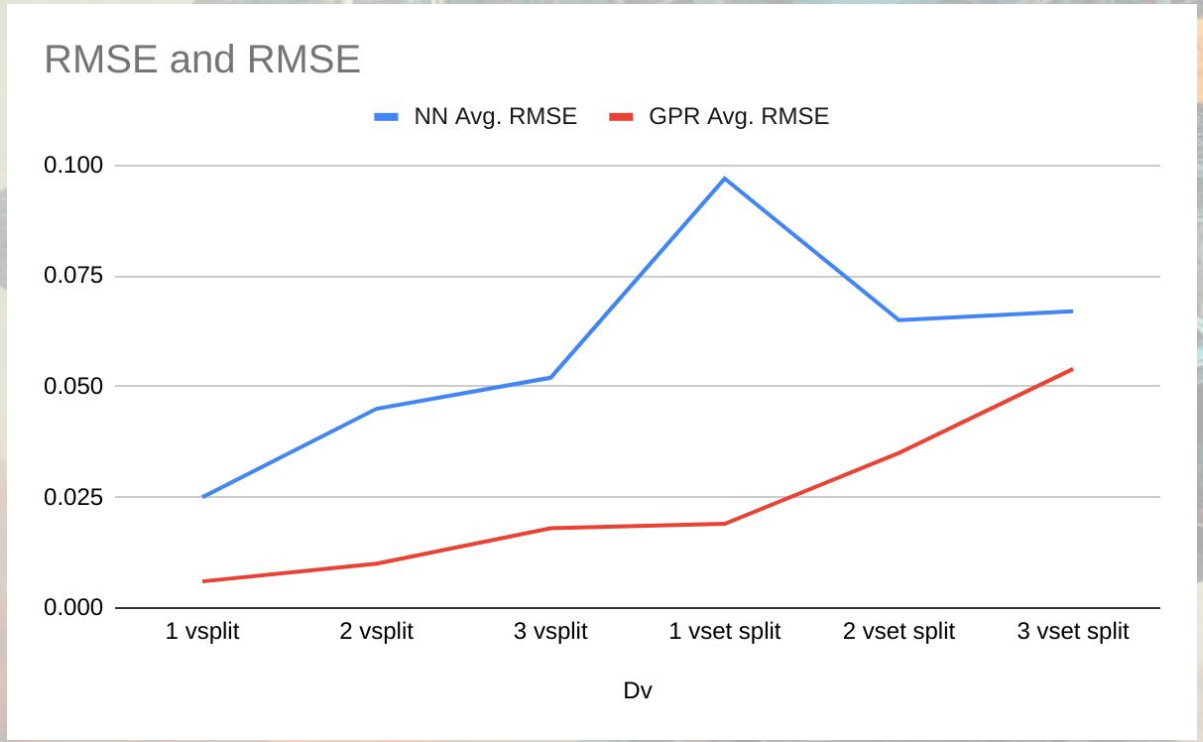
N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients



GPR $\Delta v = 1$

Blue [points are the predicted ones, while the green points are the training set

N_2-H_2 Inelastic Collisions mixed quantum-classical rate coefficients



Preliminary new results after a deeper grid search of better hyperparameters

NN [64; 64; 64] batch 10 epochs 100

GPT Mattern Kernel v = 2

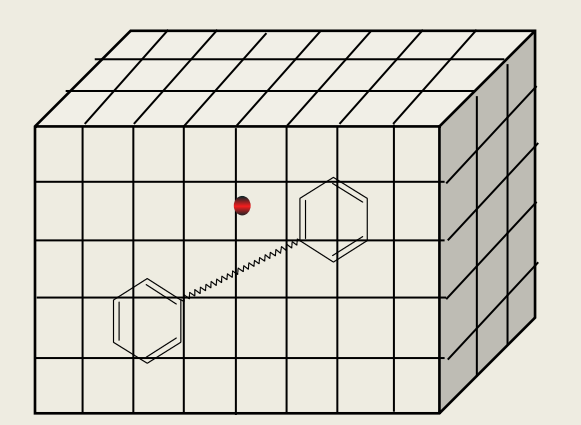
- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

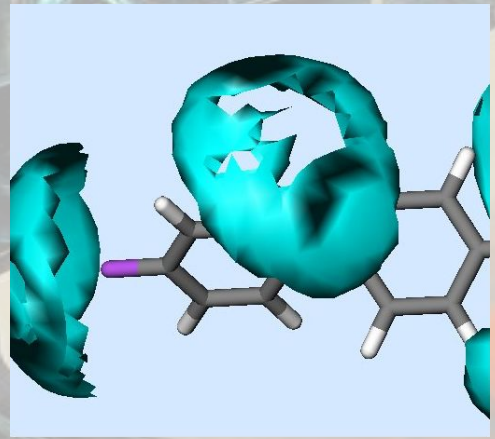
Machine Learning and the GRID Force-Fields

- **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} that is the product of three terms terms:

- E_r based on the distance between the target and the probe given in kcal/mol
- The other two, both ranging in the interval 0–1. They are dimensionless functions of the angles t and p made by the hydrogen bond (HB) at the target and the probe atoms respectively

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

Emin → dEmin

Er assumes relative values in case of interaction with a HB acceptor or donor complementary probe and is parametrized by two values: **Emin is the strongest hydrogen-bond attraction energy at the optimum position (Emin),** and half of the straight-line distance between donor and acceptor atom pairs which corresponds to the strongest hydrogen-bond attraction energy (Rmin).

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

The dataset is made of 66463 drug-like molecules

- We used GAMESS-US B3LYP/SVP (necessity of having a versatile basis set and method) to compute the Electrostatic Potential (EP) for each atom
- EP is converted to the so called dEmin value using linear equation derived so that for each AT all the resulting dEmin values always fall within an acceptable range

$$dEmin_{BH} = m_{BH} * EP + q_{BH}.$$

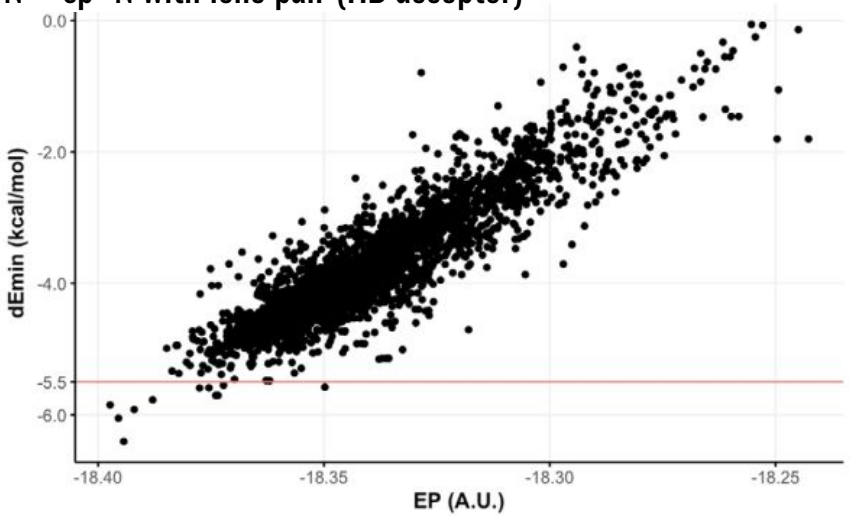
$$dEmin_{AH} = -m_{AH} * EP - q_{AH}.$$

22 equations, each one for each AtomType

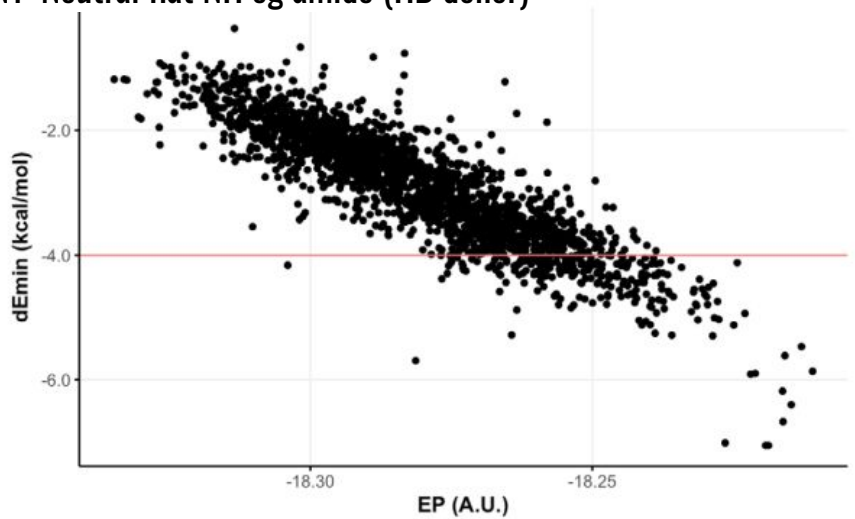
The dEmin is our label

Machine Learning and the GRID Force-Fields

N:= sp^2 N with lone pair (HB acceptor)



N1 Neutral flat NH eg amide (HB donor)



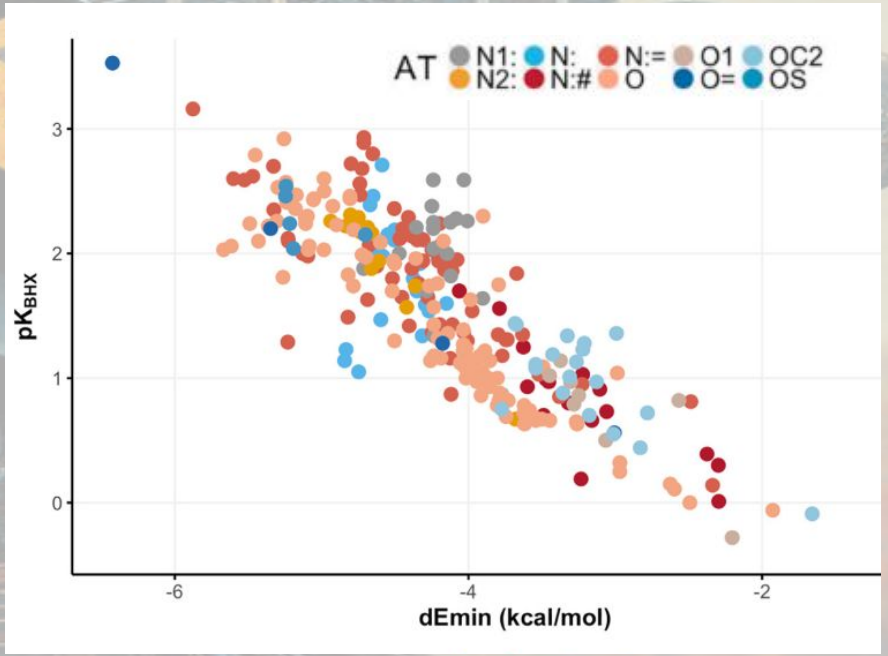
The red lines represent values of the traditional, static Emin of the GRID force field, namely -5.5 for N:= and -4.0 for N1 atom types. dEmin, dynamic Emin

Machine Learning and the GRID Force-Fields

Does chemically sound to use the dEmin in the the E_{HB} term ?

We decided to test the correlation of the proposed dEmin to those experimental hydrogen-bonding (HB) properties.

dEmin versus H-bond basicity scale for the Kenny dataset (279 atoms, R – Pearson = 0.85).

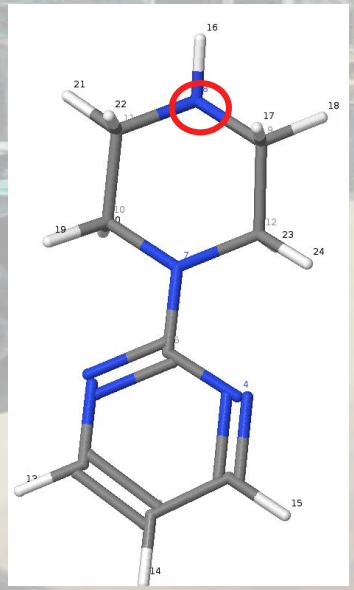


Machine Learning and the GRID Force-Fields

We have a good label, now we need to select the feature (descriptor) to use in the model

The molecular environment is described by a tree-structured molecular fingerprint with a length of 10 bond distances

```
0   1    8 N_3H 122
1   2    9 C.3 326   11 C.3 326
2   2   12 C.3 629   10 C.3 629
3   1    7 N.3_ar 1016
4   1    5 C.ar+ 1250
5   2    4 NPYM 1706    6 NPYM 1706
6   2    3 C.ar+ 1856    1 C.ar+ 1856
```



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}

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

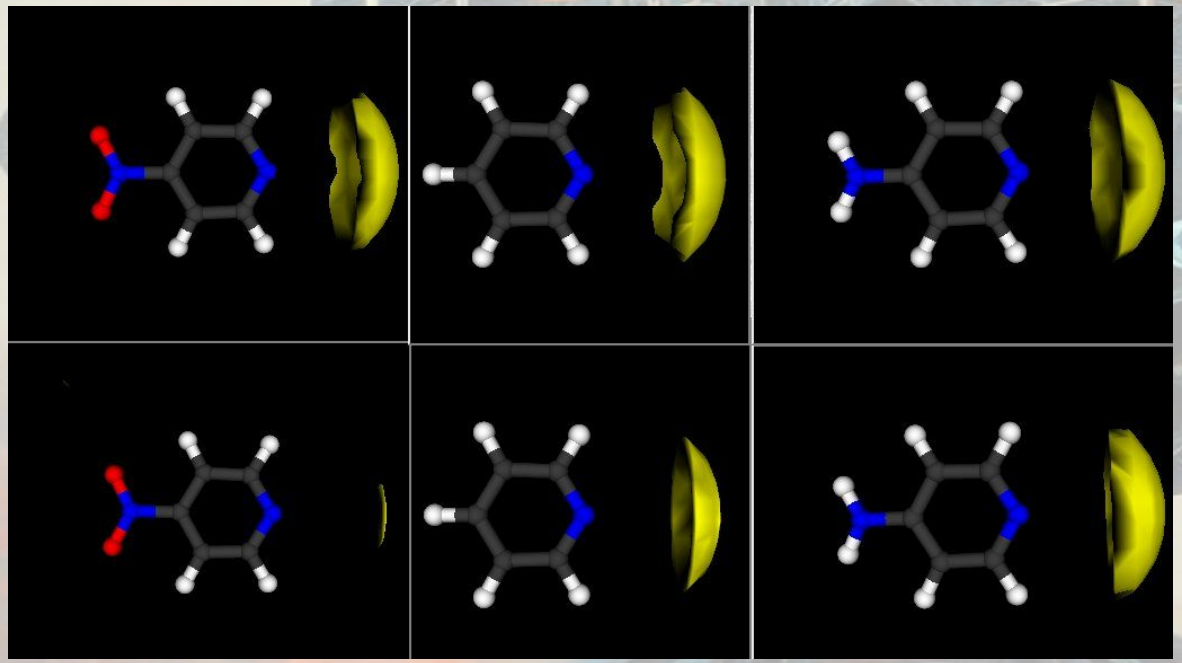
Machine Learning and the GRID Force-Fields

Using this approach, 22 PLS models were built relating atomic environment to dEmin for the HB GRID atom types (some of the models results are reported validated using leave-one-out crossvalidation)

AT	Description	H-bond type	Atoms	LV	R^2	Q^2	SDEC (kcal/Mol)	SDEP (kcal/Mol)
N:	sp3 (tertiary) nitrogen, accepting one H-bond	A	6954	9	0.92	0.88	0.56	0.41
N1:	sp3 (secondary) nitrogen, donating one hydrogen and accepting one H-bond	A	3941	8	0.91	0.84	0.24	0.49
		D	4776	7	0.96	0.92	0.30	0.53
N2:	sp3 (primary)nitrogen, donating up to two hydrogen and accepting one H-bond	A	3618	8	0.84	0.71	0.26	0.38
		D	4895	7	0.95	0.92	0.30	0.41
ON	oxygen of nitro or nitroso group, accepting up to two H-bond	A	4907	8	0.82	0.69	0.26	0.38
N:=	sp2 (aromatic) nitrogen, accepting one H-bond	A	27,140	12	0.91	0.89	0.35	0.47
N::	sp2 nitrogen with two lone pairs and one double bond	A	472	4	0.89	0.59	0.23	0.12
N:#	sp nitrogen	A	15,798	10	0.72	0.66	0.29	0.32

Machine Learning and the GRID Force-Fields

More chemically aware force-field



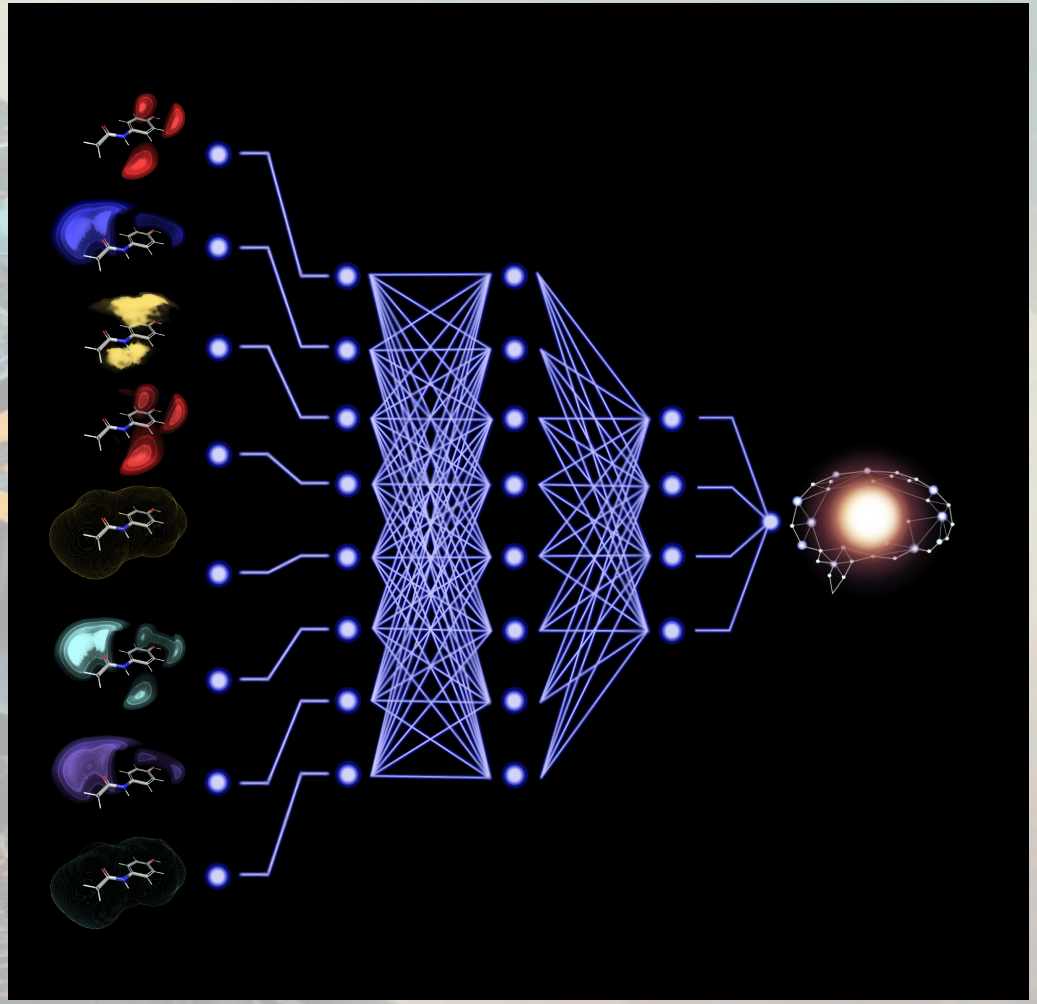
The energy values of the isocontour surfaces chosen for H-bond donating probe (“N1,” probe) was 4.0 kcal/Mol

- ML Introduction
- Deep Learning
  - Fitting using ML
    - GRID MIFs
    - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

DeepGRID

Two ingredients are needed:

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

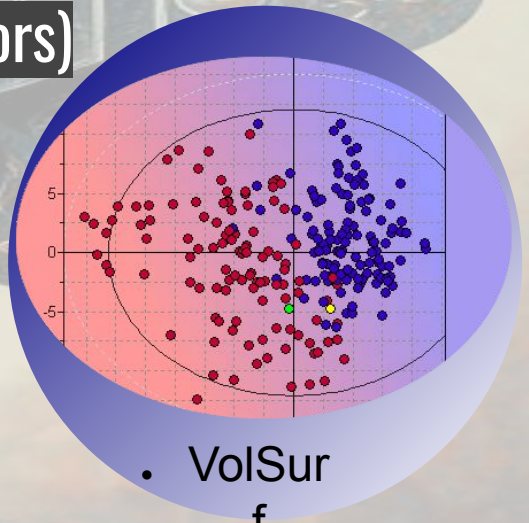


Loriano Storchi, Gabriele Cruciani, Simon Cross, "DeepGRID: Deep Learning using GRID descriptors for BBB prediction", Journal of Chemical Information and Modeling, DOI: 10.1021/acs.jcim.3c00768 (2023)


DATASET AND LABEL

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-IgBB-416** dataset A subset of the 2105 dataset which had experimental logBB values
- **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
-

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


FEATURES

DeepGRID Approach

GRAID descriptors calculated (normalised GRID MIFs, 8 channels)

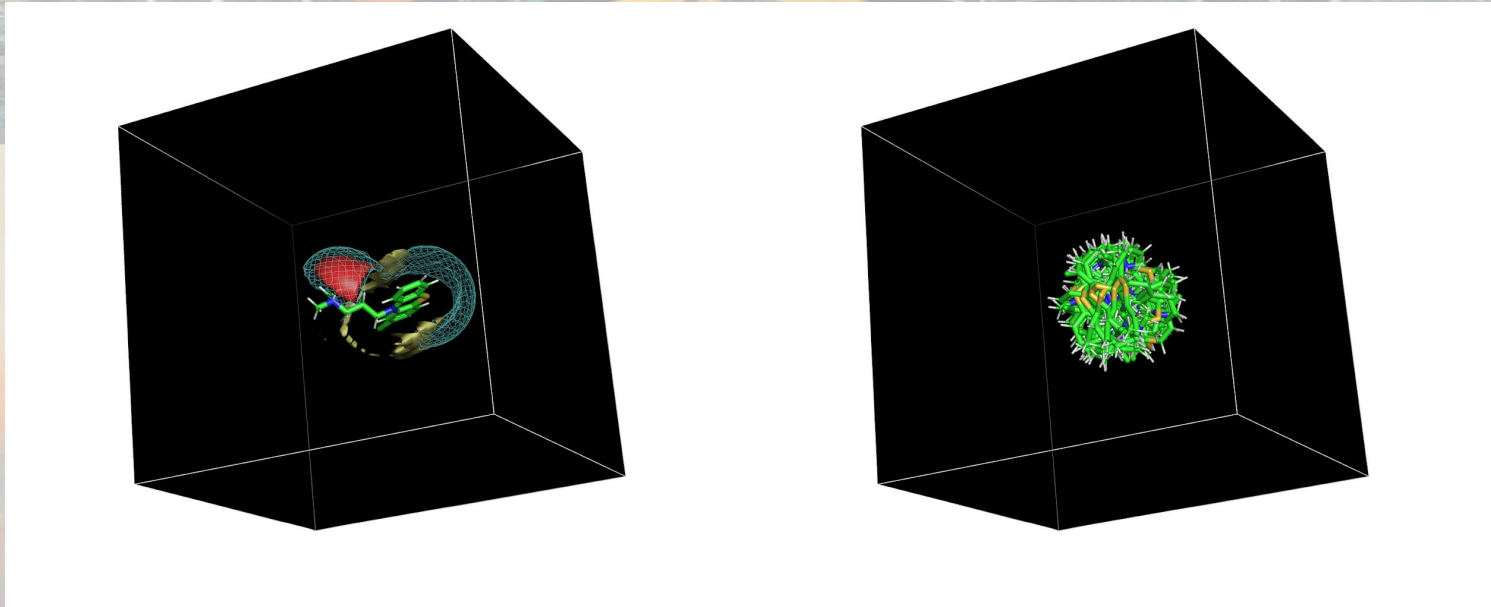
Descriptors fed into a Deep Learning CNN model

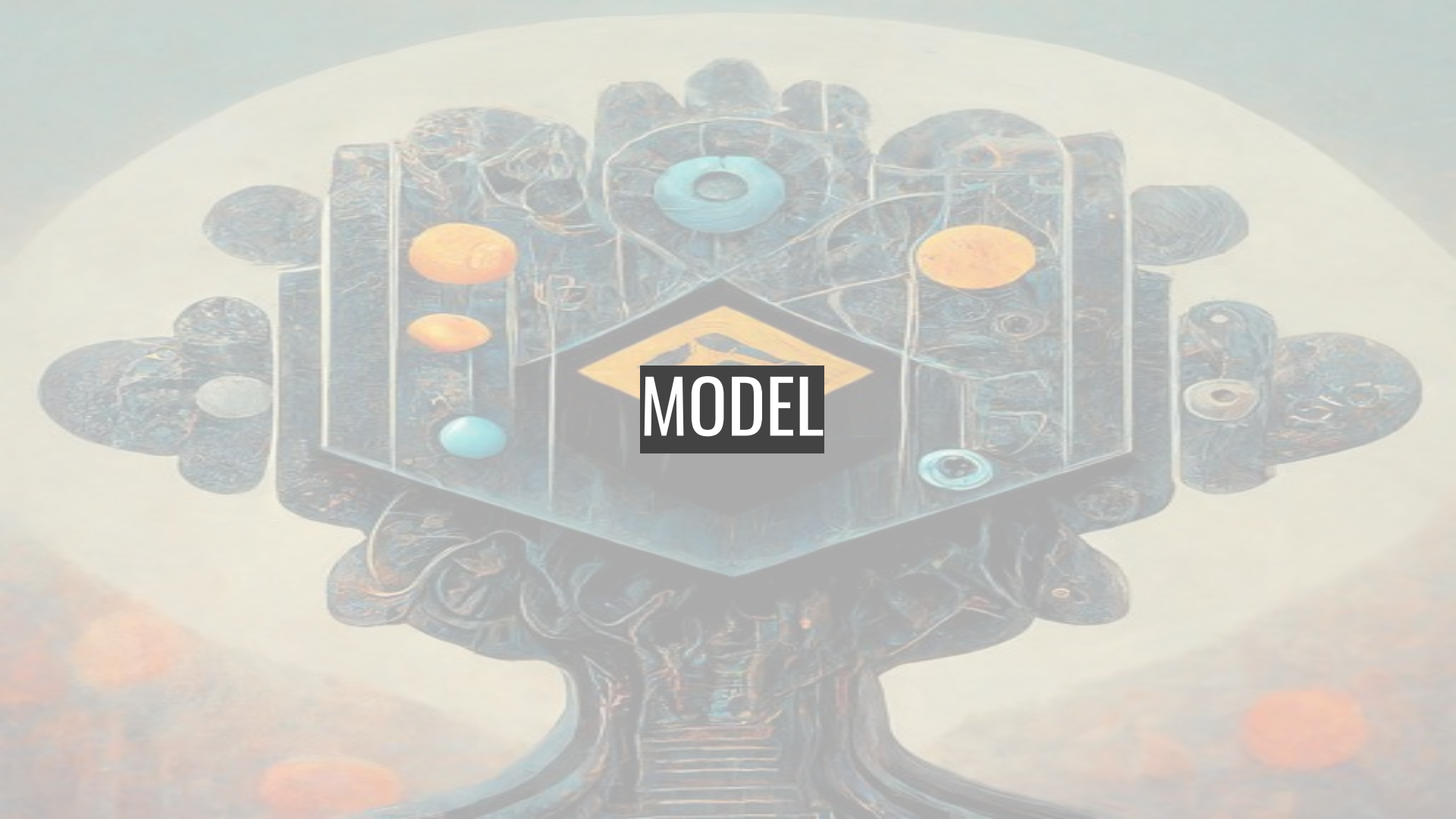
Note: in this case the training and validation sets were mixed so that different viewpoints of the same molecule were in training/validation, to allow the model to learn from the viewpoints

DeepGRID is alignment independent

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

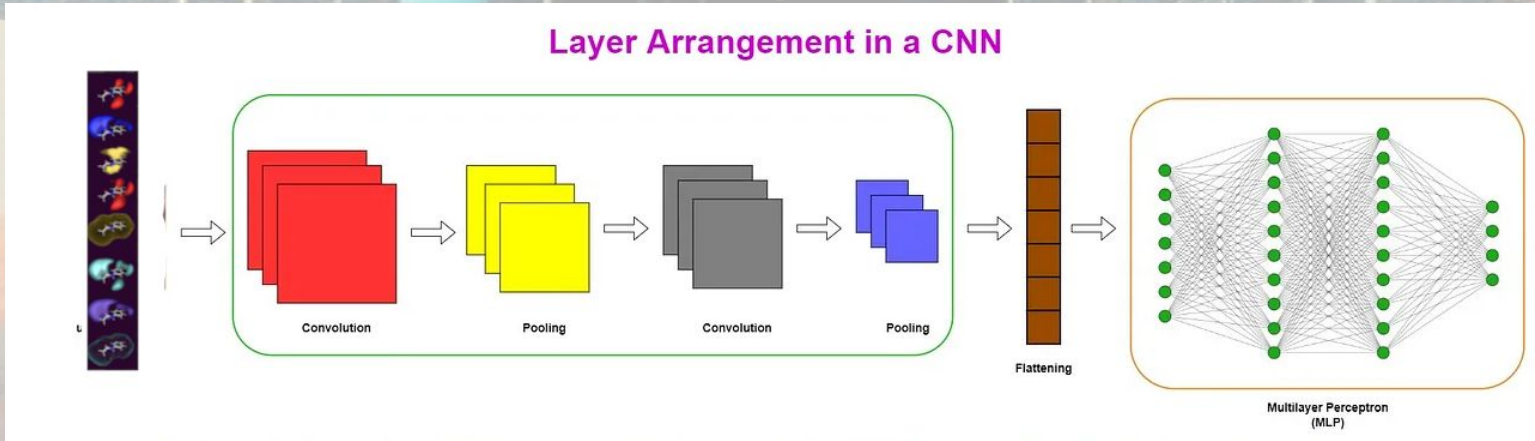
27 ‘Viewpoints’ generated by rotating the molecule around each axis

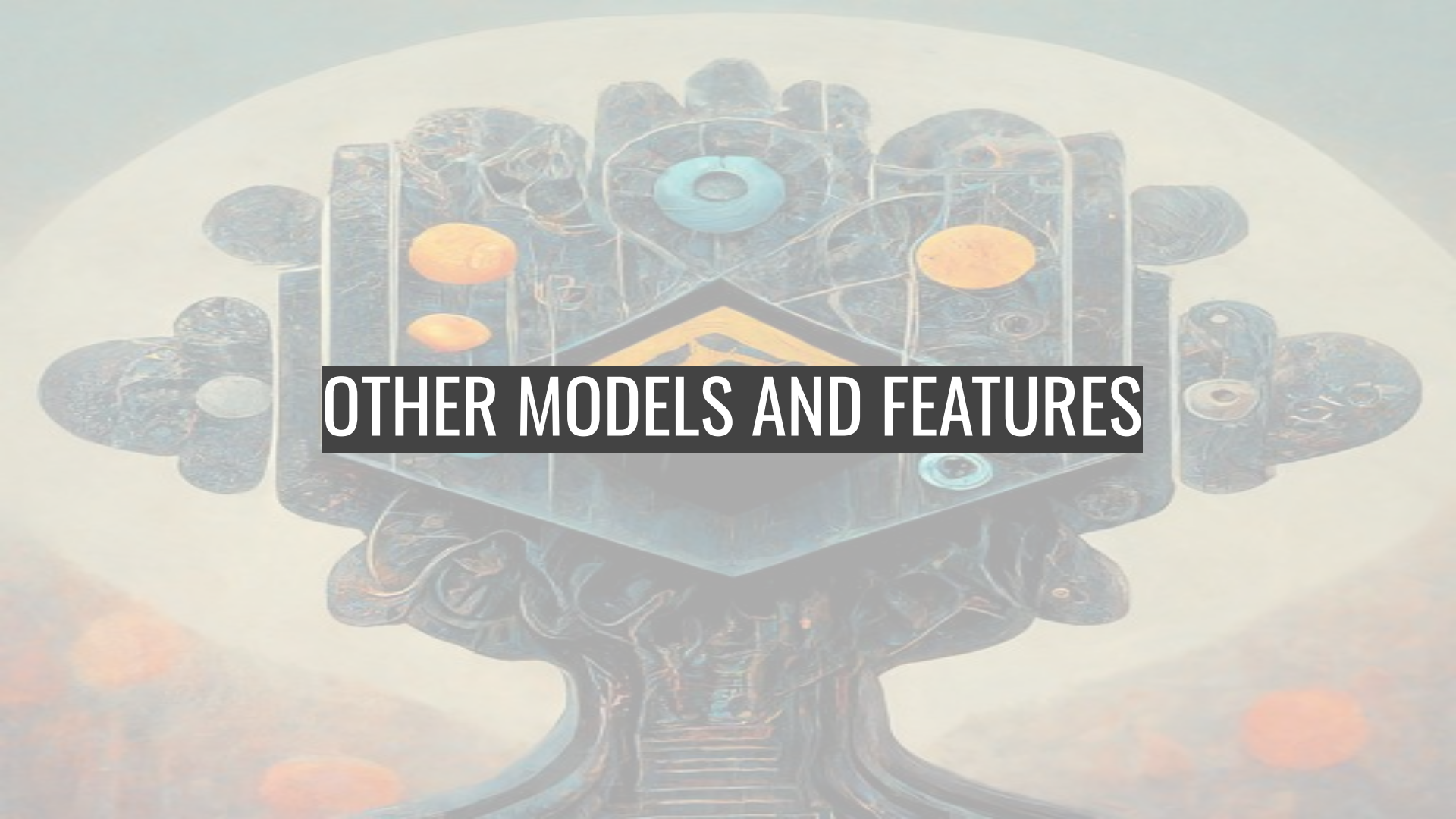



MODEL

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

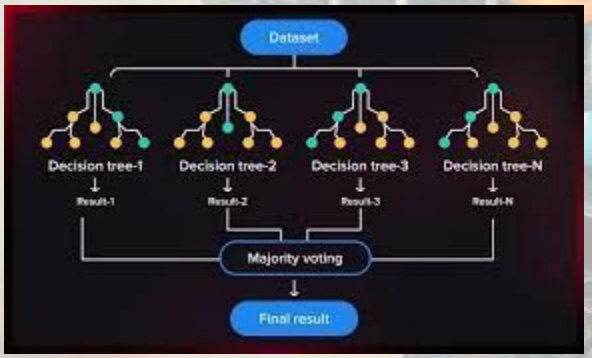
DeepGRID Hyperparameters optimization

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

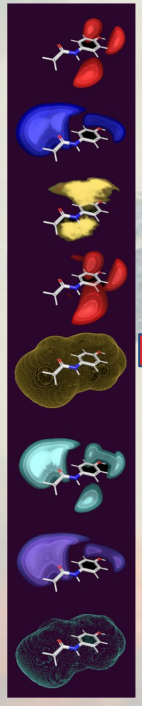
Partial Least Squares Approach

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

It is a linear relation but instead of the pure X variables we are using LV (Latent Variables) similar 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



Quite some time was needed to develop the VS descriptors

Volsurf3 Descriptors

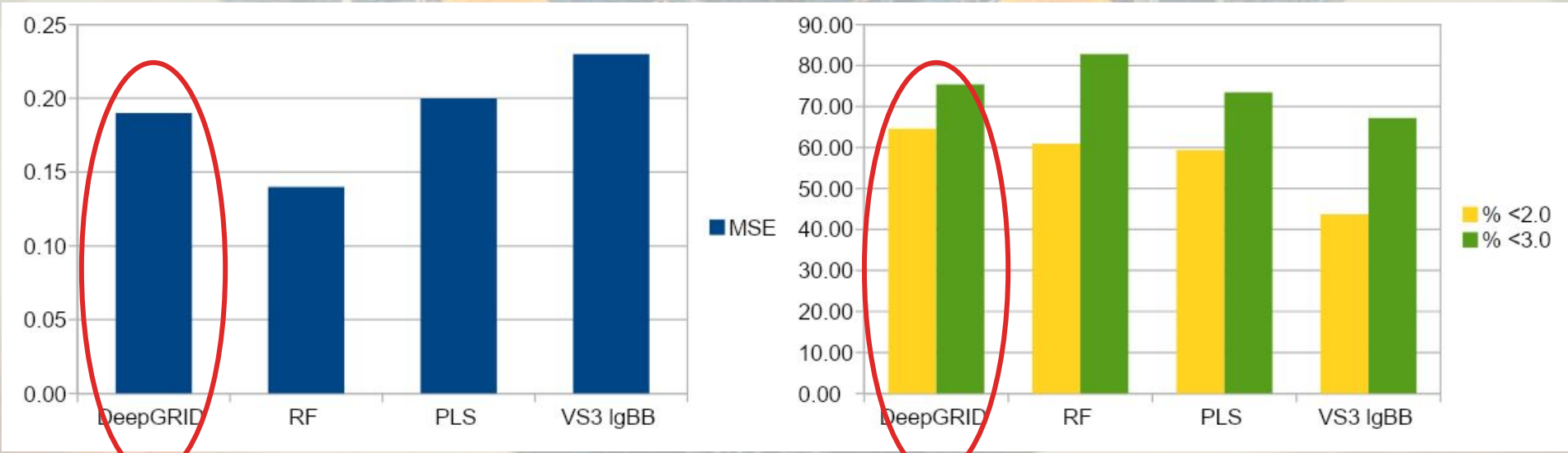
Descriptors	Probes*			Description
	OH2	DRY	O	
V	X			Molecular volume
S	X			Molecular surface
POL				Polarizability
MW				Molar mass
HB1–HB8			X	Hydrogen bonding
A				Amphiphilic moment
BV	X		X	Best volumes
W1–W8	X			Hydrophilic regions
ID1–ID8		X		Hydrophobic integy moment
Cw1–Cw8	X			Capacity factor
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


RESULTS

VS-IgBB-332 Dataset

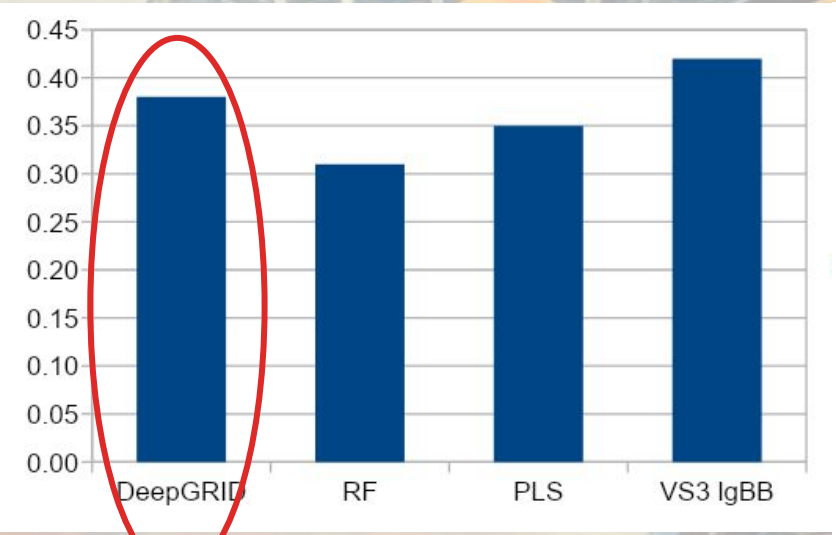


Lower is better

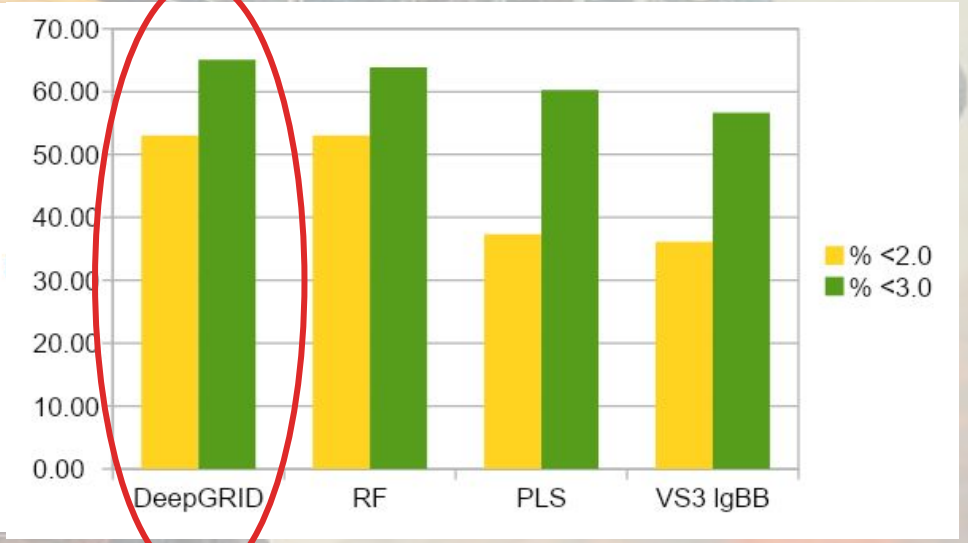
Higher is better

Light-IgBB-416 dataset is more diverse

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



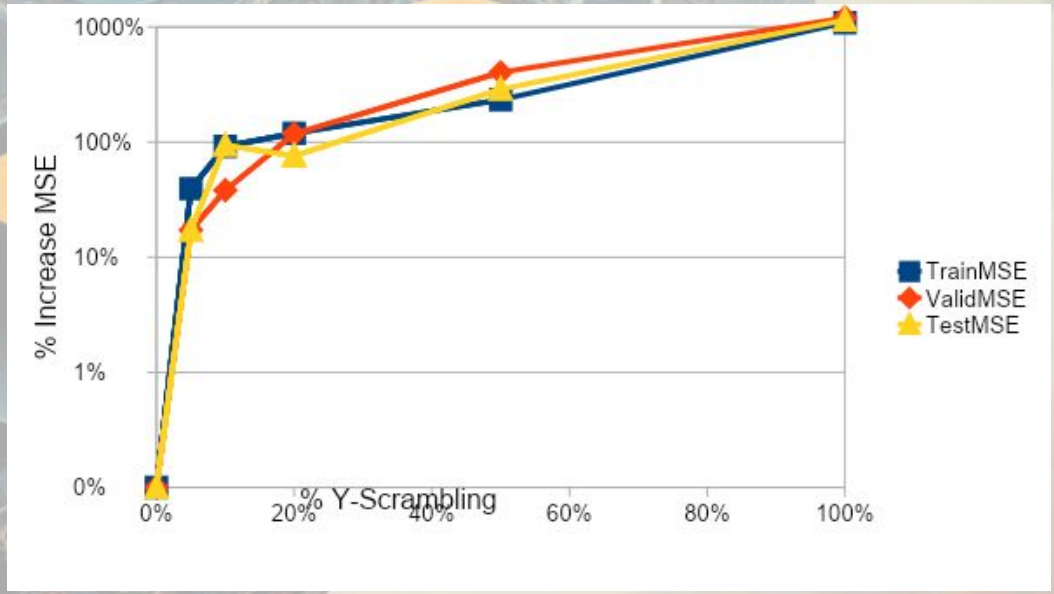
Lower is better



Higher is better

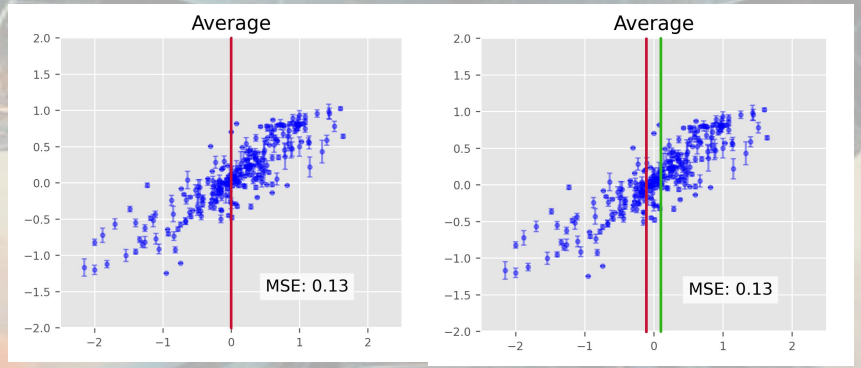
DeepGRID gives a robust model

- 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



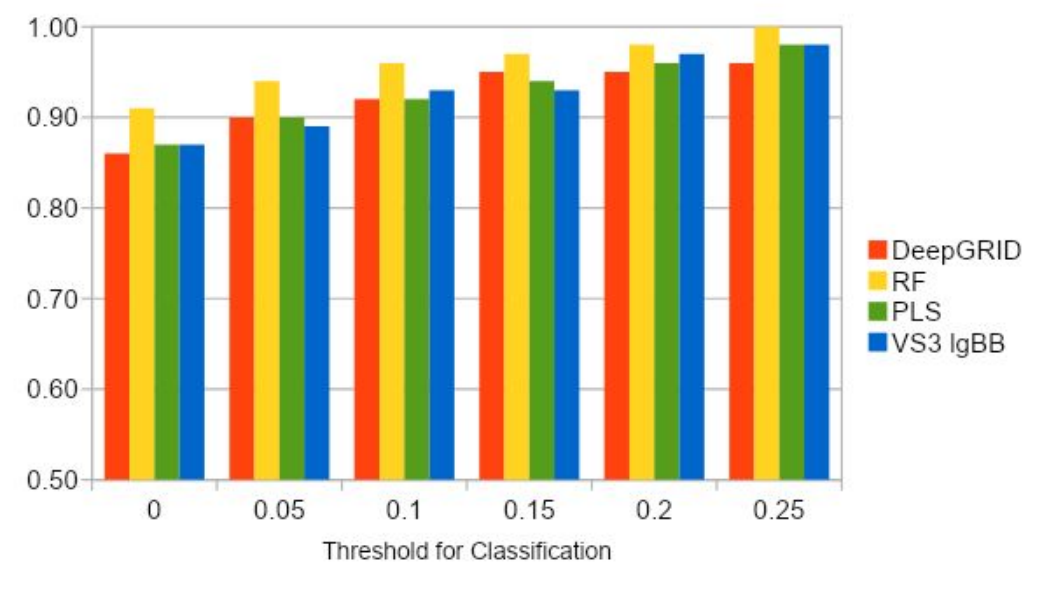
Regression → Classification

- 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**



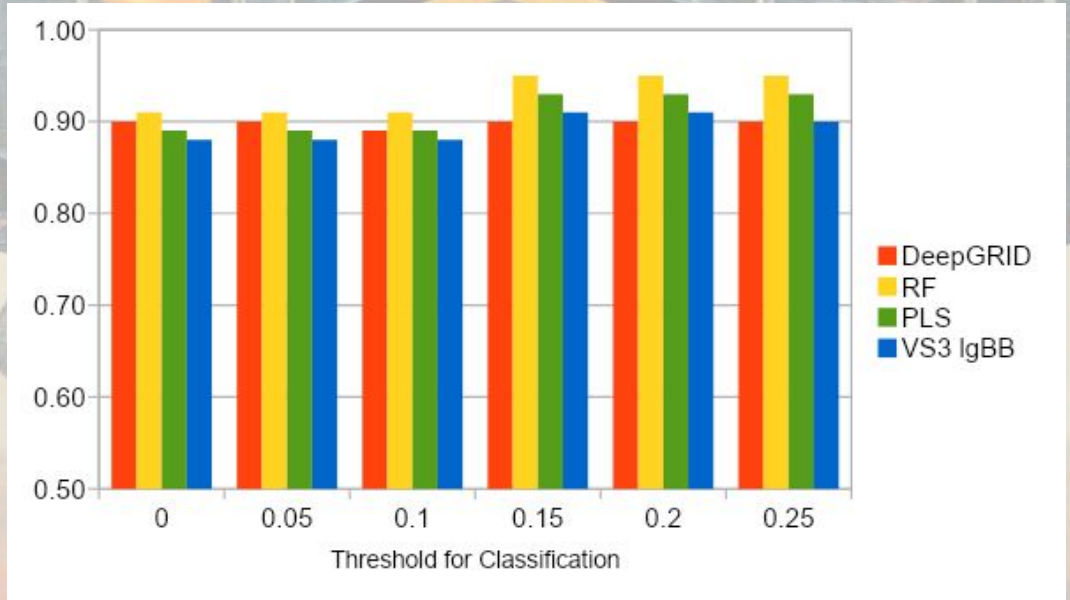
Classification: VS-IgBB-332 model

- 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

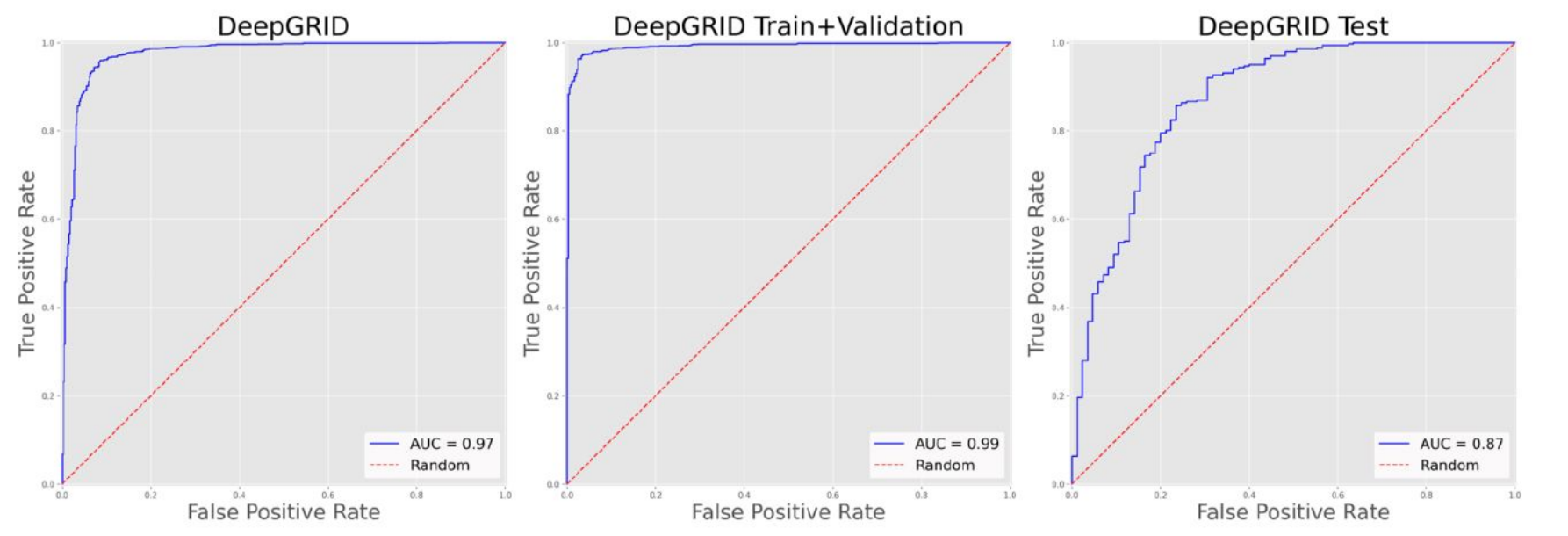


Classification Models - Light-IgBB-2105 dataset

- 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

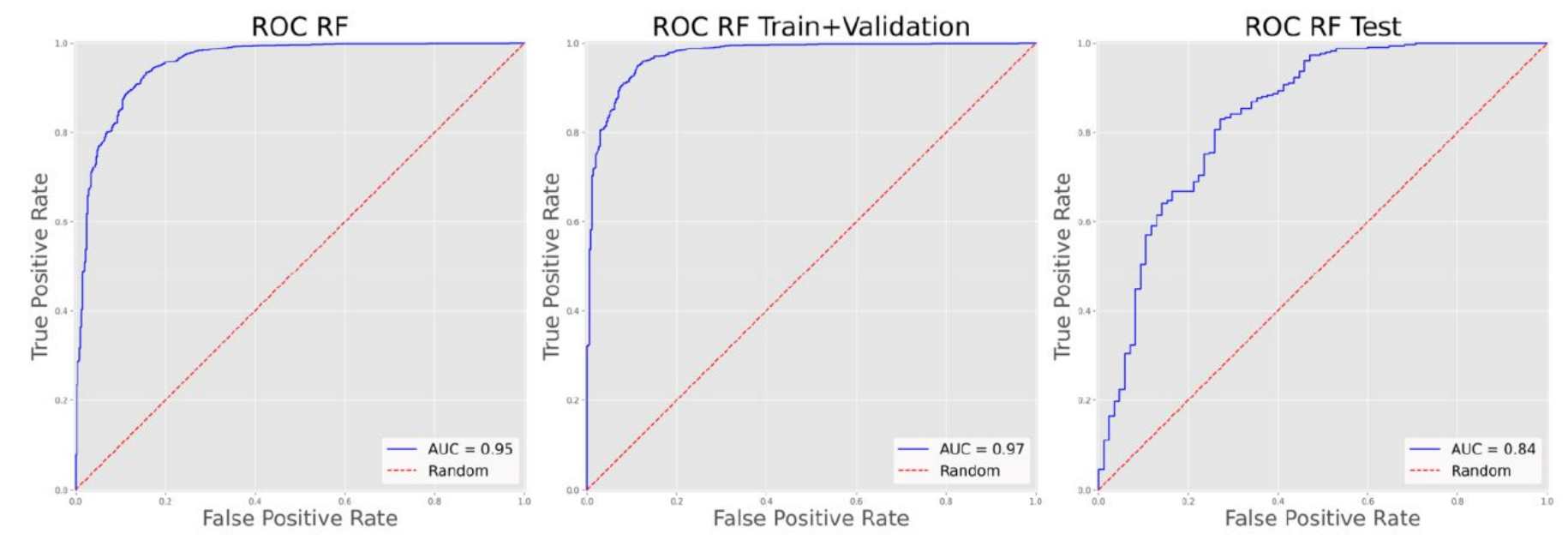
DeepGRID Classification Models - Light-IgBB-2105 dataset

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



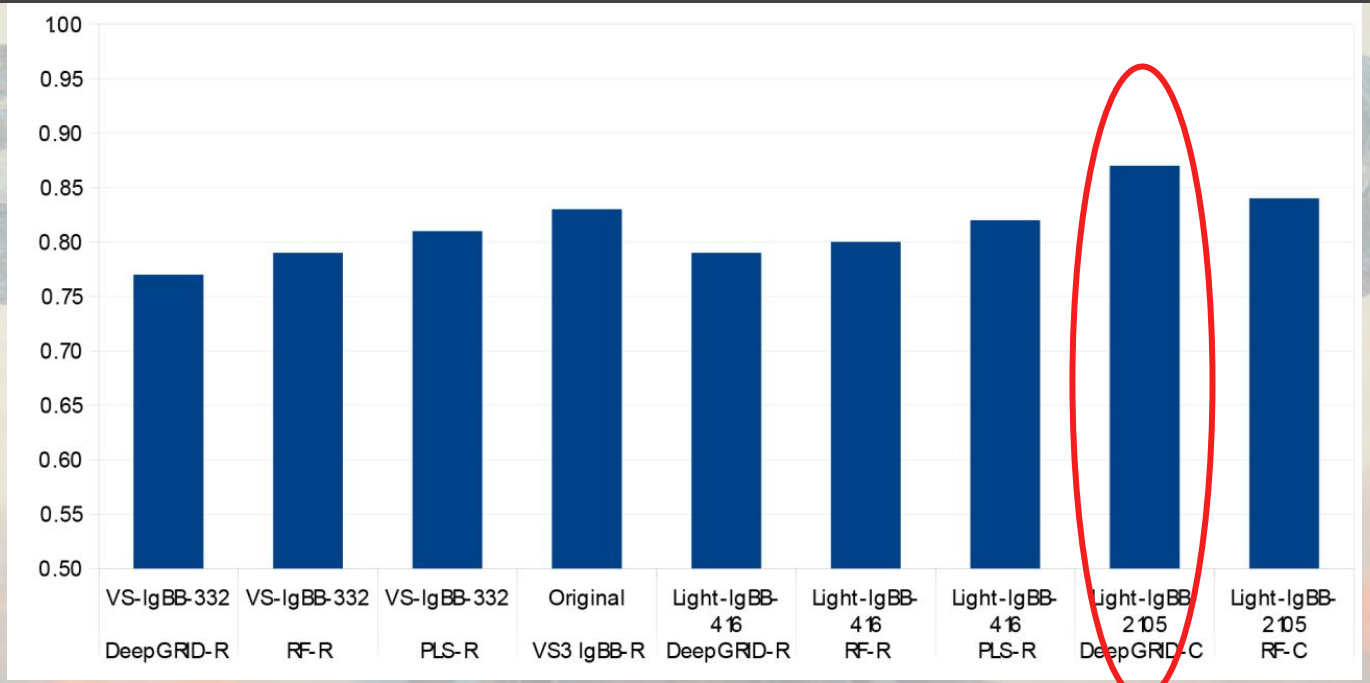
RF Classification Models - Light-IgBB-2105 dataset

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



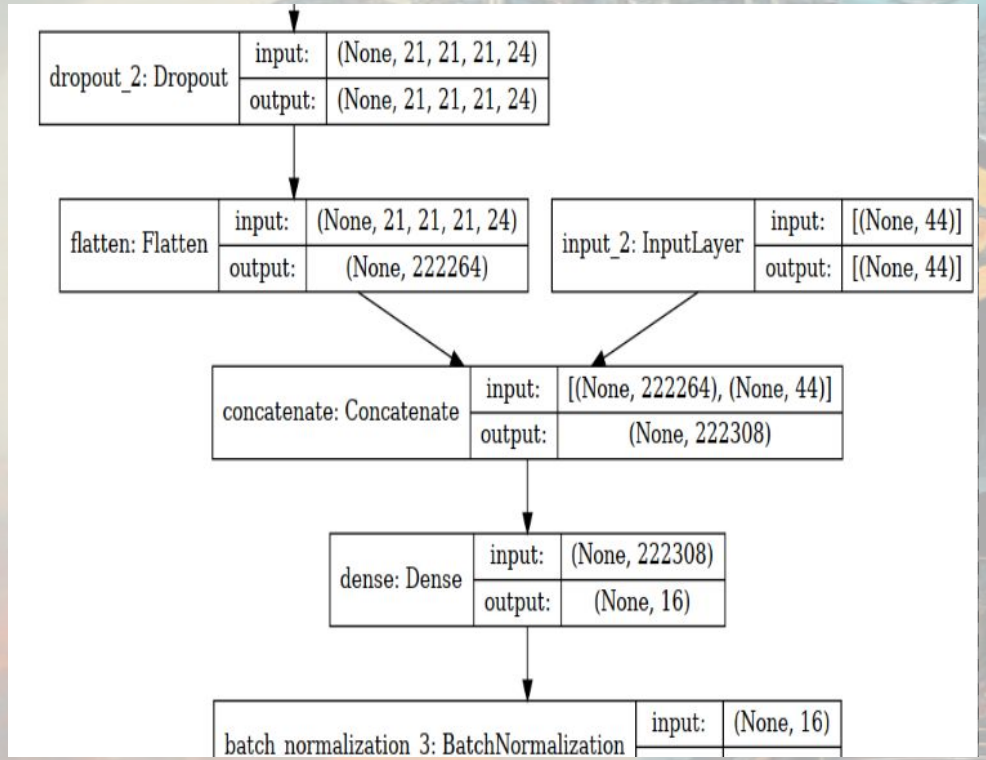
DeepGRID: mixing VS descriptors and MIF

Clearance mechanism classification fro drugs two classes :

- Metabolic Clearance: This is the most complex mechanism, involving the biotransformation of drugs into more hydrophilic metabolites to facilitate excretion.(643 compounds)
- Renal Clearance: This mechanism involves the direct excretion of drugs in the urine, typically for small, hydrophilic compounds. (329 compounds)

I am using augmentation techniques

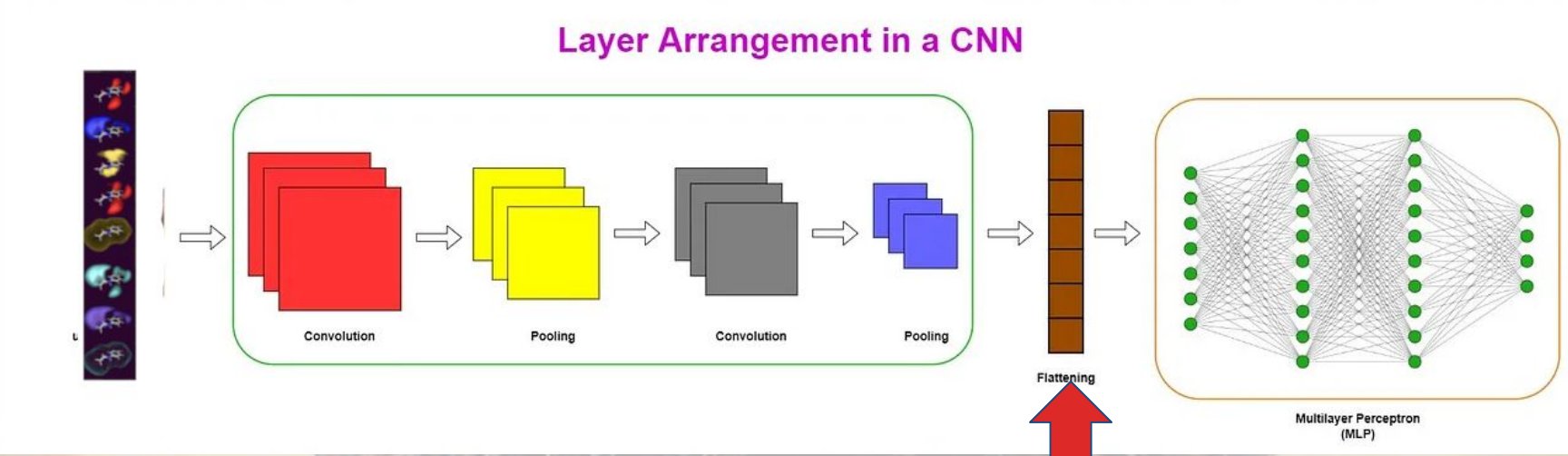
DeepGRID: mixing VS descriptors and MIF



Two model are cocananated :

- Model 1 is the CNN model
- Model 2 is a simple input layer thata is getting the VS descriptors just before the flattening layer

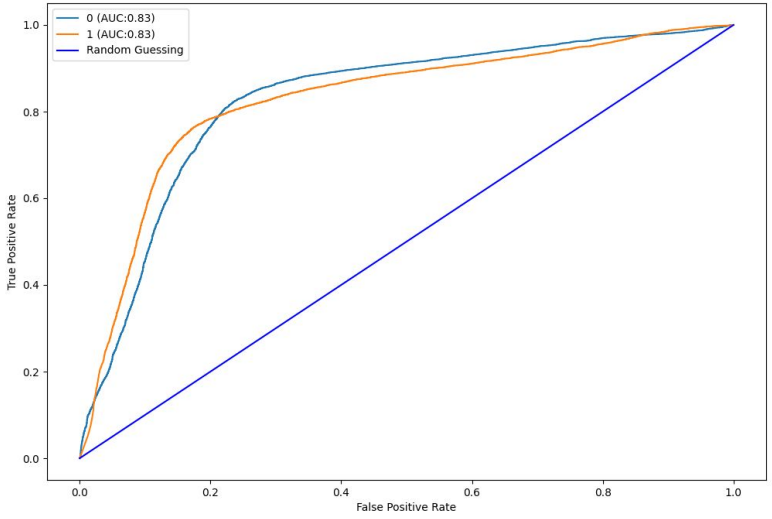
DeepGRID: mixing VS descriptors and MIF

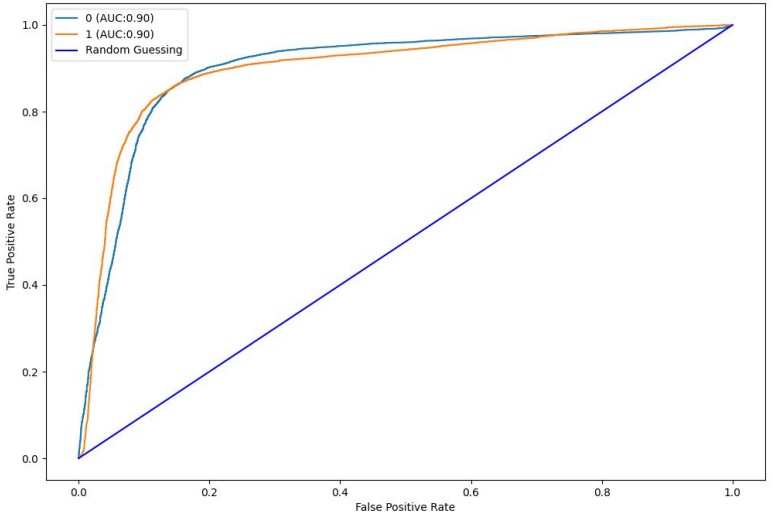


VS descriptors are appended together with the output of the Convolutional layers in the flatten layer

DeepGRID: mixing VS descriptors and MIF

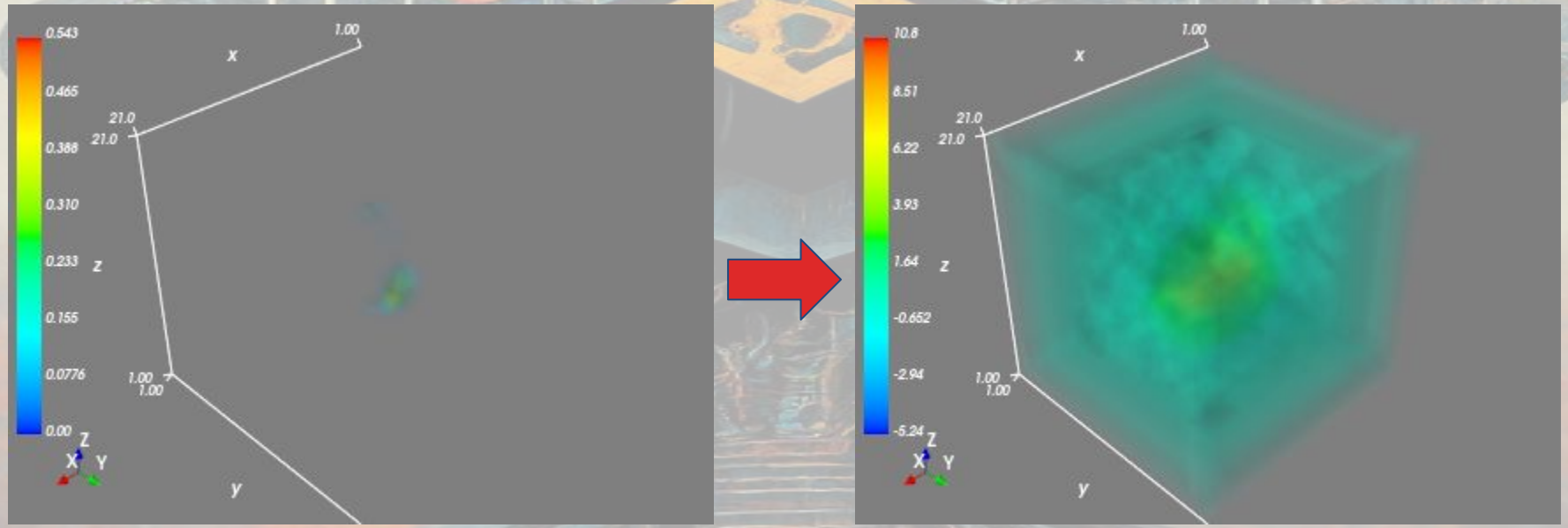
AUC Test Set without VS: **0.83 With VS : 0.90**





DeepGRID: try to understand how the CNN works

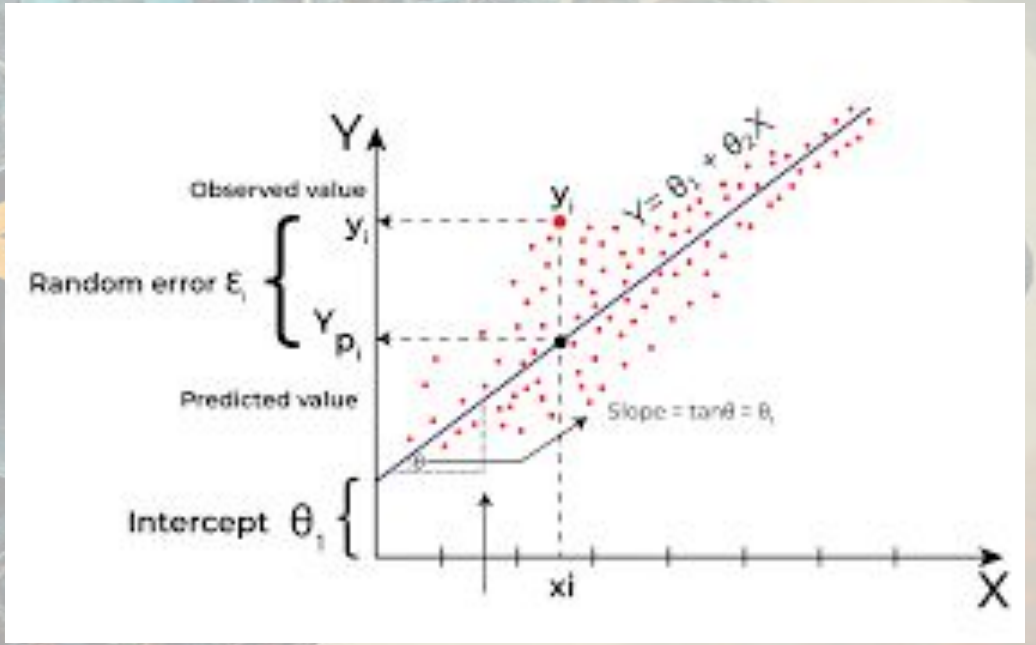
It is possible, although quite tricky, to dump the features as extracted by the Convolutional layers:



- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

Linear Regression

Linear Regression models predict a dependent variable (Y) based on independent variables (X).
The relationship between the variables is assumed to be linear.
Models are relatively simple and easy to interpret.
Common applications include predicting sales, energy consumption, and other continuous values.
A key assumption is that the errors are normally distributed.



- ML Introduction
- Deep Learning
  - Fitting using ML
    - GRID MIFs
    - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

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 [rocksalt\ and ZB; (zinc blende) from that optimization, a classification of the most stable crystal structure semiconductor AB binary compounds **(full dataset is made of 82 compounds)**

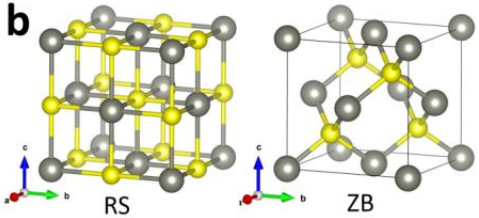
Udaykumar Gajera, Loriano Storchi, Danila Amoroso, Francesco Delodovici, Silvia Picozzi "Towards machine learning for microscopic mechanisms:a formula search for crystal structure stability based on atomic properties" Journal of Applied Physics, DOI: 10.1063/5.0088177 (2022)

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

Binary system based on Mulliken's electronegativity(EN)

A,B where EN(A) < EN(B)

2 types of atoms

7 APs per atom

IP, EA, HOMO, LUMO, r_s, r_p, r_d

5 prototype functions ($f(X)$)

$X, X^2, X^3, \sqrt{X}, e^X$

Material Feature Generator

EX-1: Gen3

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

For each possible model, calculate average of RMSEs for random N test-train splits using Linear Regression

$$\Delta E = m \times \frac{f_1(AP_1) \pm f_2(AP_2)}{f_3(AP_3) \pm f_4(AP_4)} + q$$

Selection of the best 10 formulas having the smallest RMSE for formula optimization

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

Top formulas to predict ΔE and further physical analysis

(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_p)

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

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

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

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

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
  - Fitting using ML
    - GRID MIFs
    - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

A Scoring Function

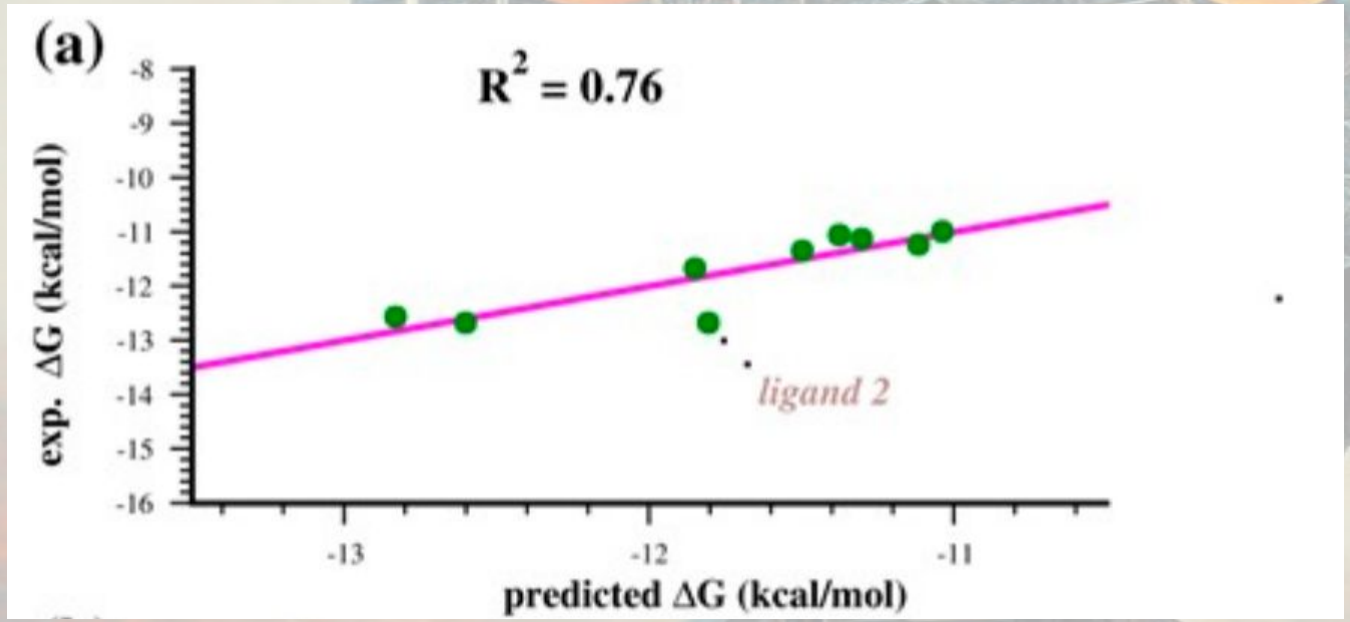
Predicting ligand-metalloenzyme binding affinity, focusing on human Carbonic Anhydrase II (hCA II) inhibitors. It combines fragment molecular orbital (FMO) and GRID approaches,

- FMO Calculations: FMO2 calculations were performed on reduced ligand-receptor complexes to assess binding energies and pair interaction energies.
- GRID Calculations: GRID was used to calculate hydrophobic interaction fields and quantify hydrophobic interactions.
- Dataset: A set of benzenesulfonamide ligands of hCA II was selected as a case study.

Roberto Paciotti, Nazzareno Re,Loriano Storchi, "Combining the Fragment Molecular Orbital and GRID Approaches for the Prediction of Ligandâ€“Metalloenzyme Binding Affinity: The Case Study of hCA II Inhibitors", Molecules, DOI: 10.3390/molecules29153600 (2024)

A Scoring Function

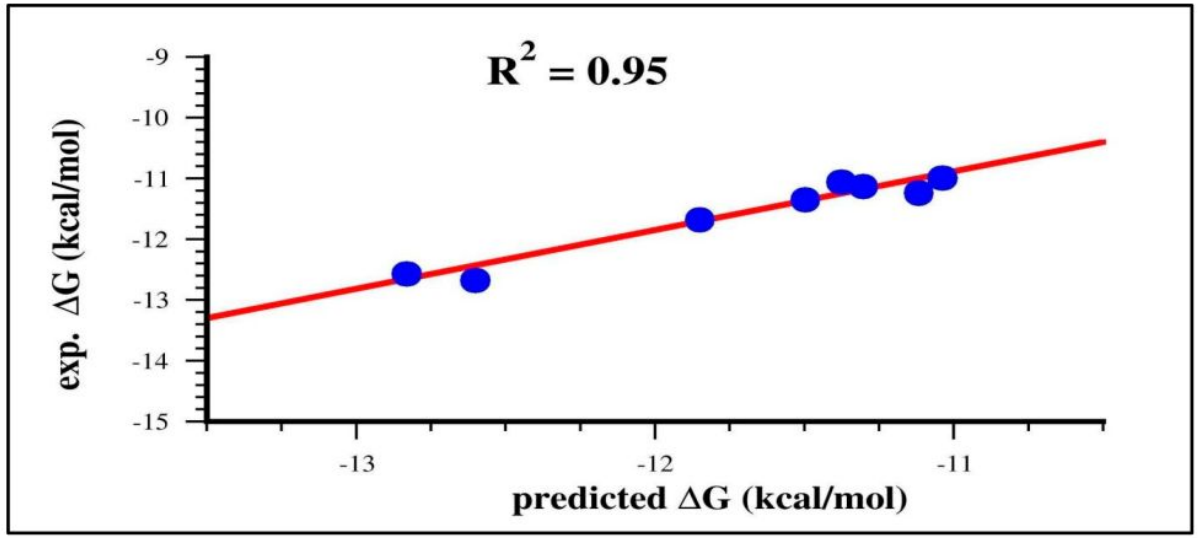
$$\Delta G = -7.4\{[0.7(\log P)^3 - 0.5(e^{\text{HIE-E}})]/[0.5(\text{F2LE})^3 - 0.4(\text{HIE-E})^5]\} - 13$$



A portion of the Ligand 2 structure connected to the benzenesulfonamide is polar compared to other ligands, which determines, in principle, a better interaction with water molecules. Thus, we hypothesize that the its binding pose in the experimental conditions assumed in the measurement of the Ki could be influenced by surrounding water molecules and be slightly different from that observed in the crystal structure.

A Scoring Function

$$\Delta G = -7.4\{[0.7(\log P)^3 - 0.5(e^{HIE\text{-}E})]/[0.5(F2LE)^3 - 0.4(HIE\text{-}E)^5]\} - 13$$



To improve the binding affinity of the benzenesulfonamide there should be a certain balance between electrostatic and hydrophobic interactions in order to minimize the denominator and maximize maximize the binding affinity

- ML Introduction
- Deep Learning
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

Interpretable Machine Learning

Interpretable Machine Learning: Techniques to explain and understand model predictions.
Provides insights into feature importance and model decision-making process.
Helps build trust and transparency in ML systems.
Enables identification of biases and potential areas for improvement.
Enhances model debugging and validation.
Two classes of method:
- Model agnostic
- Model specific

Random Forrest and Permutation Feature Importance

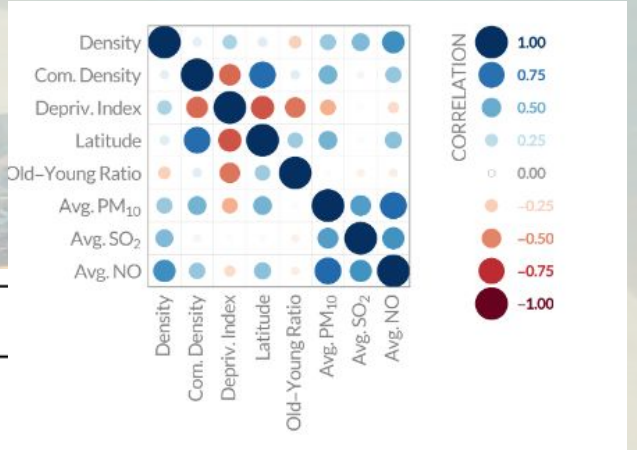
Use the RF model not for prediction purpose but to detect how much a feature is important respect to the others. Two ingredients:

- 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.

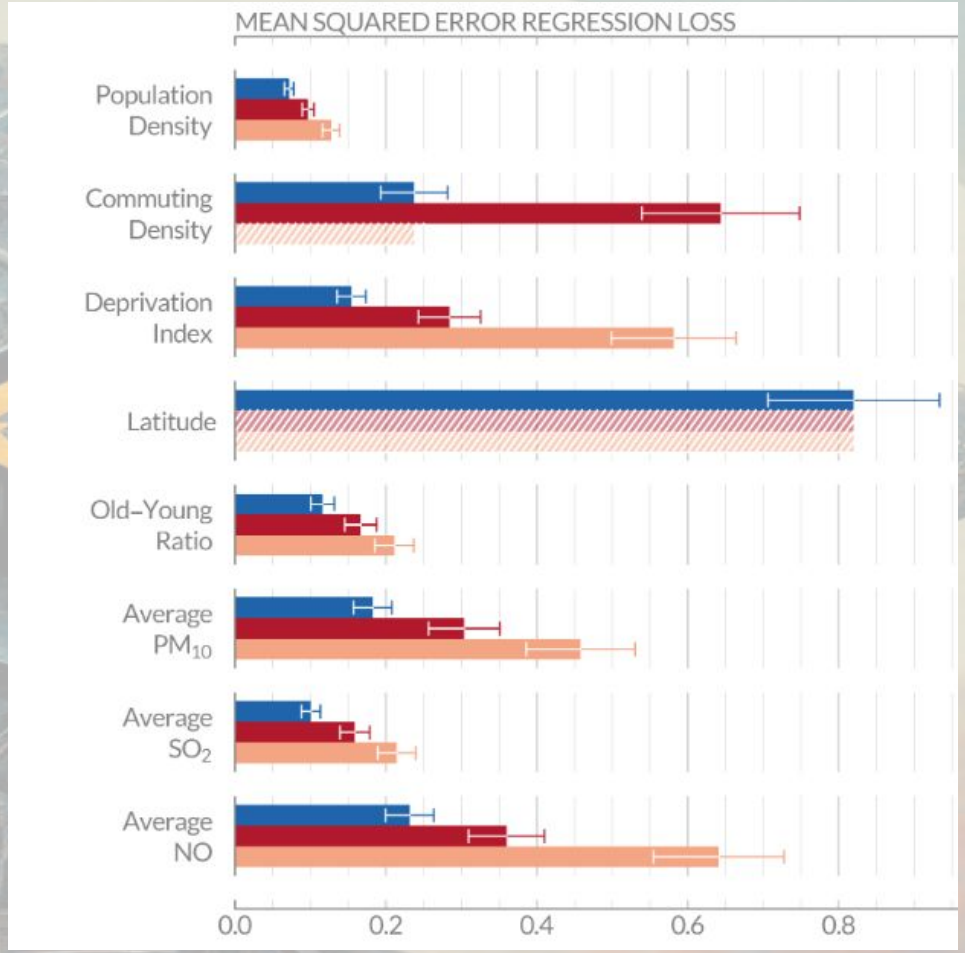


Results

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
  - Fitting using ML
  - GRID MIFs
  - DeepGRID
- Linear Regression
  - A formula search for crystal structure stability
  - A Scoring Function
- Interpretable Machine Learning
- Conclusions and future work

Conclusions and next step

- ML techniques can be used with both big and small datasets
- Different approaches (i.e. directly interpretable of more Black Box like models)
- ML can be use also to gain some insights about data rather then “simply” predict
- Features can be booth simple as well as more structured data

Formula generator using different approaches/models: PLS, PCR, but also GAN

THANK YOU

Leonardo Belpassi
Daniele Spiga
Tommaso Tedeschi
Diego Ciangottini
Mirco Tracolli
Danila Amoroso
Silvia Picozzi

Simon Cross
Sara Tortorella
Emanuele Carosati
Gabriele Cruciani
Roberto Paciotti
Cecilia Coletti
Qizhen Hong
Giovanni Bistoni

Bibliography

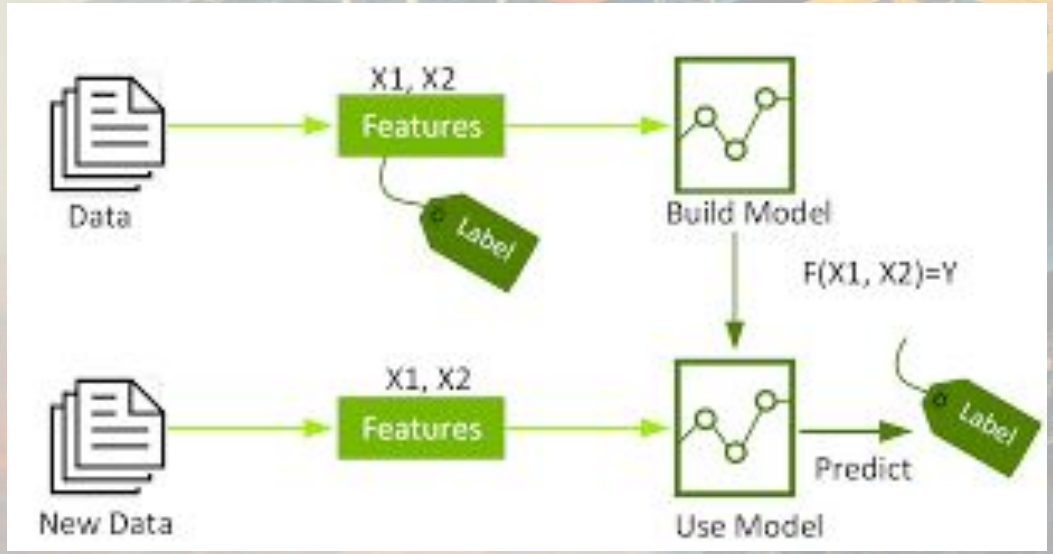
- Loriano Storchi, Gabriele Cruciani, Simon Cross, "DeepGRID: Deep Learning using GRID descriptors for BBB prediction", Journal of Chemical Information and Modeling, DOI: 10.1021/acs.jcim.3c00768 (2023)
- Qizhen Hong, Loriano Storchi, Quanhua Sun, Massimiliano Bartolomei, Fernando Pirani, Cecilia Coletti, "Improved Quantumâ^'Classical Treatment of N2â^'N2 Inelastic Collisions: Effect of the Potentials and Complete Rate Coefficient Data Sets", Journal of Chemical Theory Computation, DOI: 10.1021/acs.jctc.3c01103 (2023)
- 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)
- Qizhen Hong, Loriano Storchi, Massimiliano Bartolomei, Fernando Pirani, Quanhua Sun, Cecilia Coletti, "Inelastic N2+H2 collisions and quantum-classical rate coefficients: large datasets and machine learning predictions" The European Physical Journal D, DOI: 10.1140/epjd/s10053-023-00688-4 (2023)
- Daniele Spiga, Diego Ciangottini, Alessandro Costantini, Sara Cutini, Cristina Duma, Jacopo Gasparetto, Pasquale Lubrano, Barbara Martelli, Elisabetta Ronchieri, Davide Salomoni, Giusy Sergi, Loriano Storchi, Mirco Tracolli, "Open-source and cloud-native solutions for managing and analyzing heterogeneous and sensitive clinical Data", Proocing of Science, https://pos.sissa.it/415/022/pdf (2022)
- 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)
- 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))
- Cruciani G., Milletti F., Storchi L., Sforna G., Goracci L., "In silico pK(a) Prediction and ADME Profiling", Chemistry and Biodiversity, DOI: 10.1002/cbdv.200900153 (2009).
- F. Milletti, L. Storchi, G. Sforna, G. Cruciani, "New and original pka prediction method using of GRID molecular interaction fields", Journal of Chemical Information and Modeling, DOI: 10.1021/ci700018y (2007).
- F. Milletti, L. Storchi, G. Sforna, S. Cross, G, Cruciani, "Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases", Journal of Chemical Information and Modeling, DOI: 10.1021/ci800340j (2009).
- F. Milletti, L. Storchi, L. Goracci, S. Bendels, B. Wagner, M. Kansy, G. Cruciani, "Extending pKa prediction accuracy: high-throughput pKa measurements to understand pKa modulation of new chemical series", European Journal of Medicinal Chemistry, DOI: 10.1016/j.ejmech.2010.06.026 (2010).
- 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",submitted to Air Quality, Atmosphere & Health


BACKUP


ML INTRODUCTION

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



$$Y = F_{a,b,c}(X)$$

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

Machine Learning

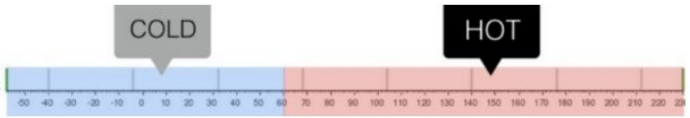
Regression

What will be the temperature tomorrow?

Fahrenheit

Classification

Will it be hot or cold tomorrow?



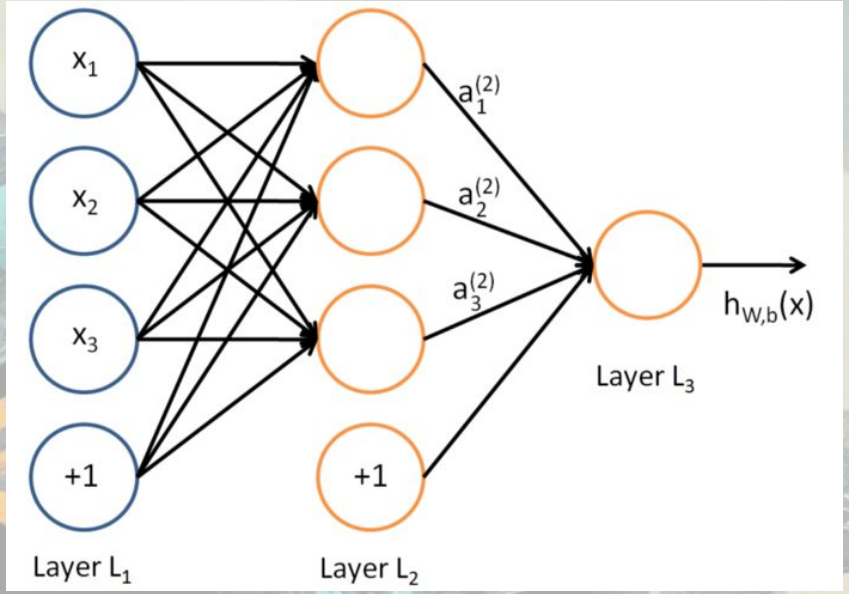
Fahrenheit

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

Label: is the temperature for the regression and

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**



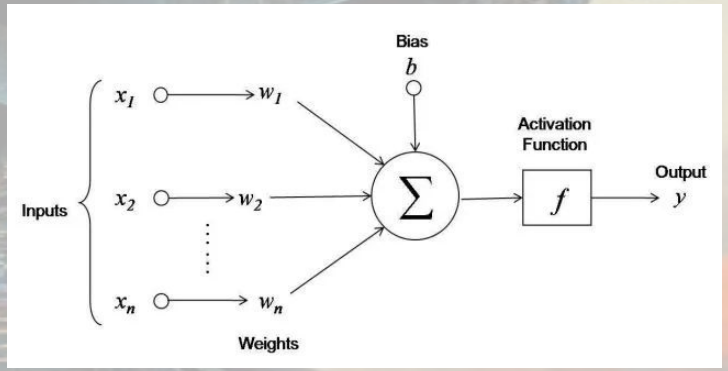
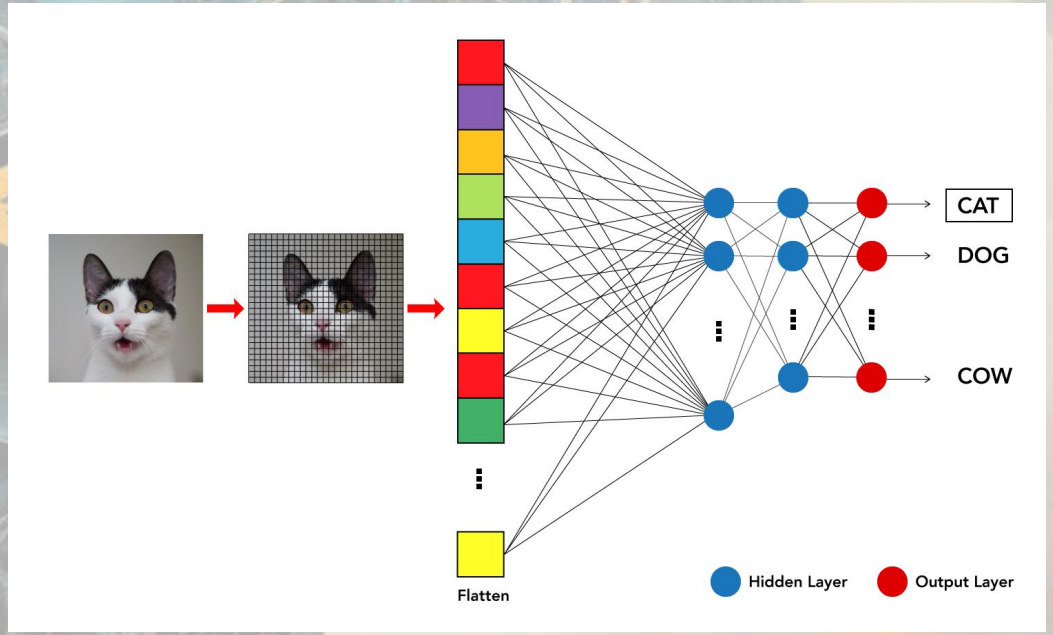


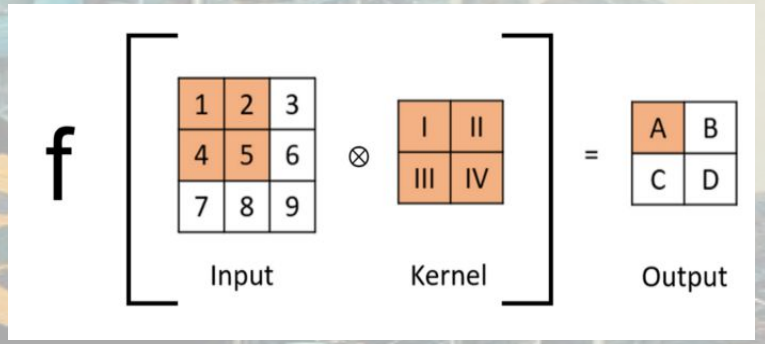
Image Recognition

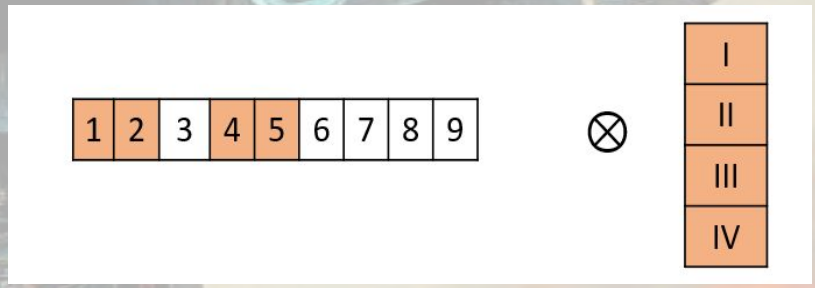
- Recognition of people, animals, objects, places etc from digital images
- Trained using thousands of pre-labelled images
- Uses the pixels in each image as descriptors
- Trained to recognise if the image shows a certain class



Convolutional Layers – extracting feature

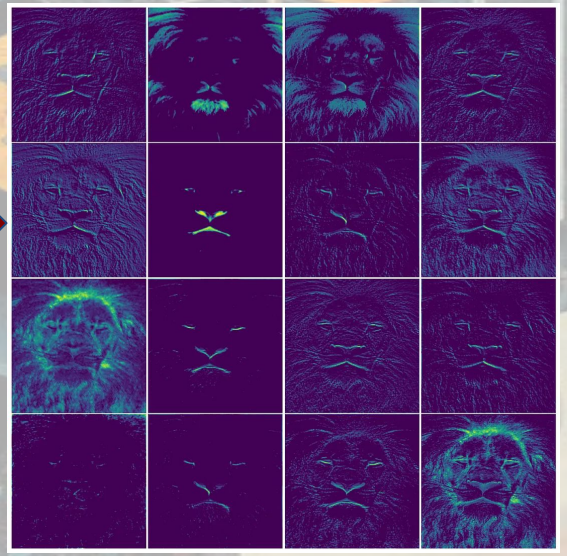
- An image is a cuboid having its length, width (dimension of the image), and height (i.e the channel 3 channelrs for RGB
- Kernel slides across the height and width of the image input and dot product of the kernel and the image are computed





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.


CURVE FITTING

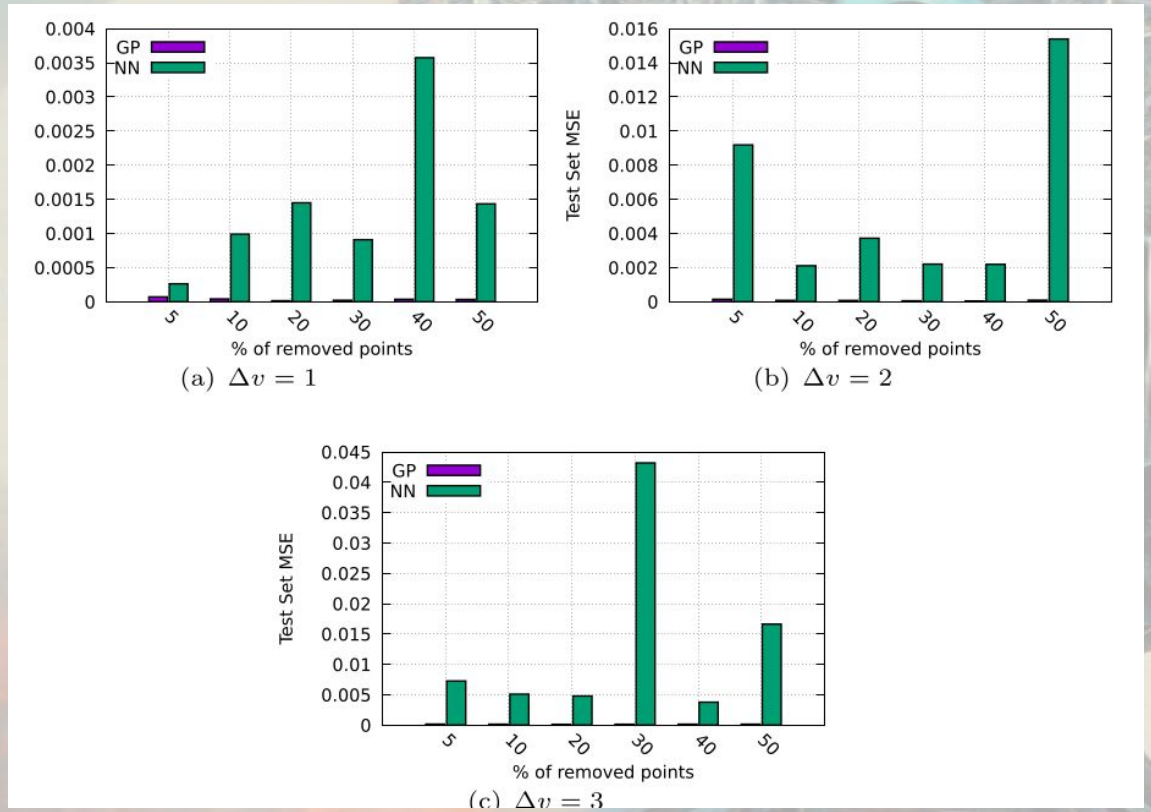
N_2-H_2 Inelastic Collisions quantum-classical rate coefficients

GPR seems to be generally the best choice in this scenario but :

Model/Scenario	Wall Time (s)
GP/Scenario1	313.392
NN/Scenario1	8.689
GP/Scenario2	420.473
NN/Scenario2	10.594
GP/Scenario3	340.595
NN/Scenario3	10.601
GP/Scenario4	69.451
NN/Scenario4	5.554

Wall times for the training processes of GP and NN models performed on an Intel(R) Xeon(R) CPU E5-1620 v4 running at 3.50GHz (i.e., without GPU)

N_2-H_2 Inelastic Collisions quantum-classical rate coefficients



Test set MSE values for the two models obtained by removing an increasing number of random points (5% to 50%) from the training set. The three panels correspond to processes (5) with $\Delta v = 1, 2, 3$, respectively


DEEP GRID

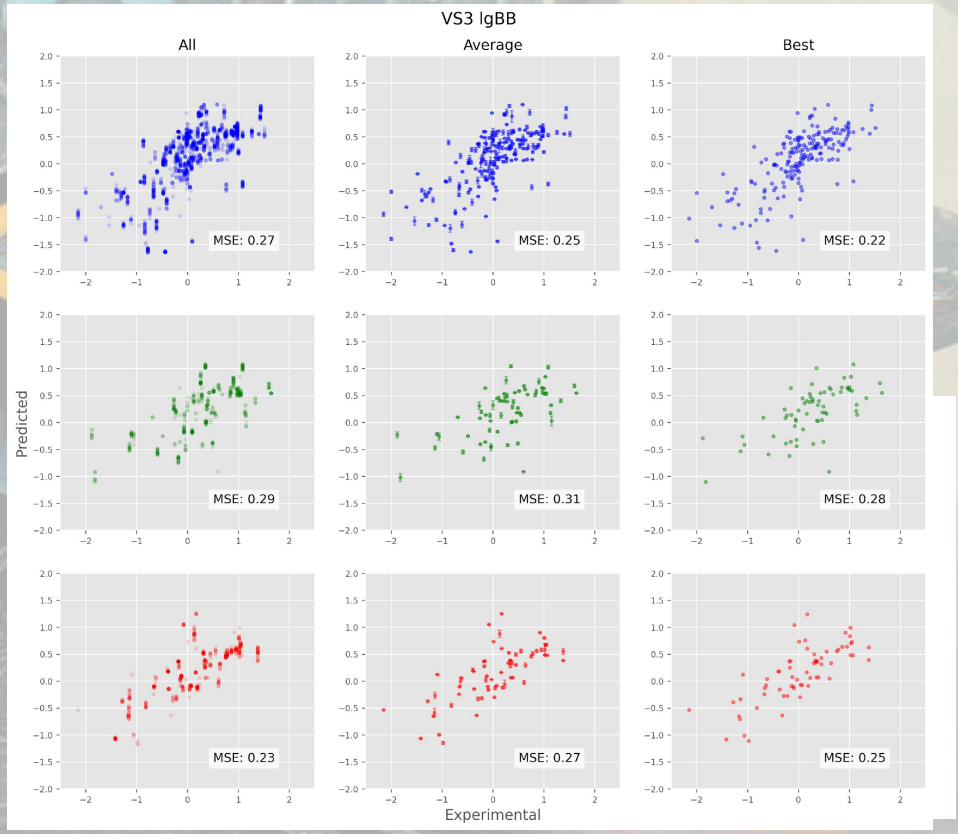
VolSurf 3 performance, VS-lgBB-332 dataset

The plots show Exp vs Pred for:

- All conformers
- Avg prediction across confs
- Best prediction by conf

The Avg prediction for the test set gives:

- MSE: 0.27



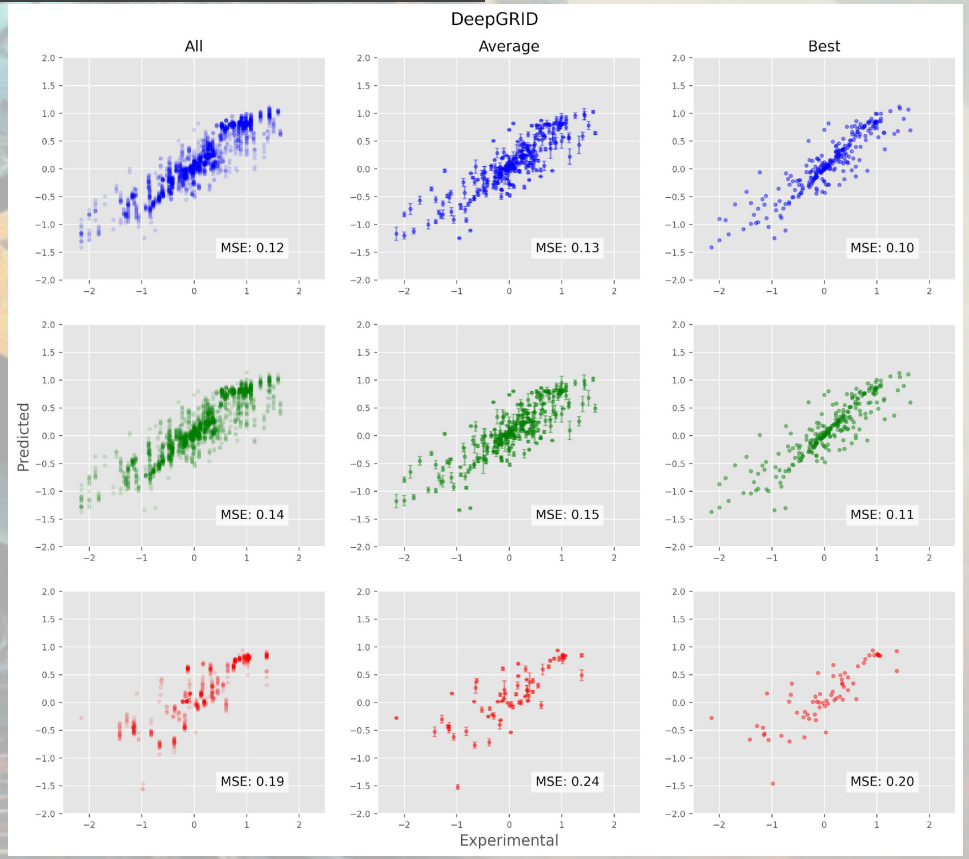
DeepGRID performance, VS-IgBB-332 dataset

The plots show Exp vs Pred for:

- All conformers
- Avg prediction across confs
- Best prediction by conf

The Avg prediction for the test set gives:

- MSE: 0.24

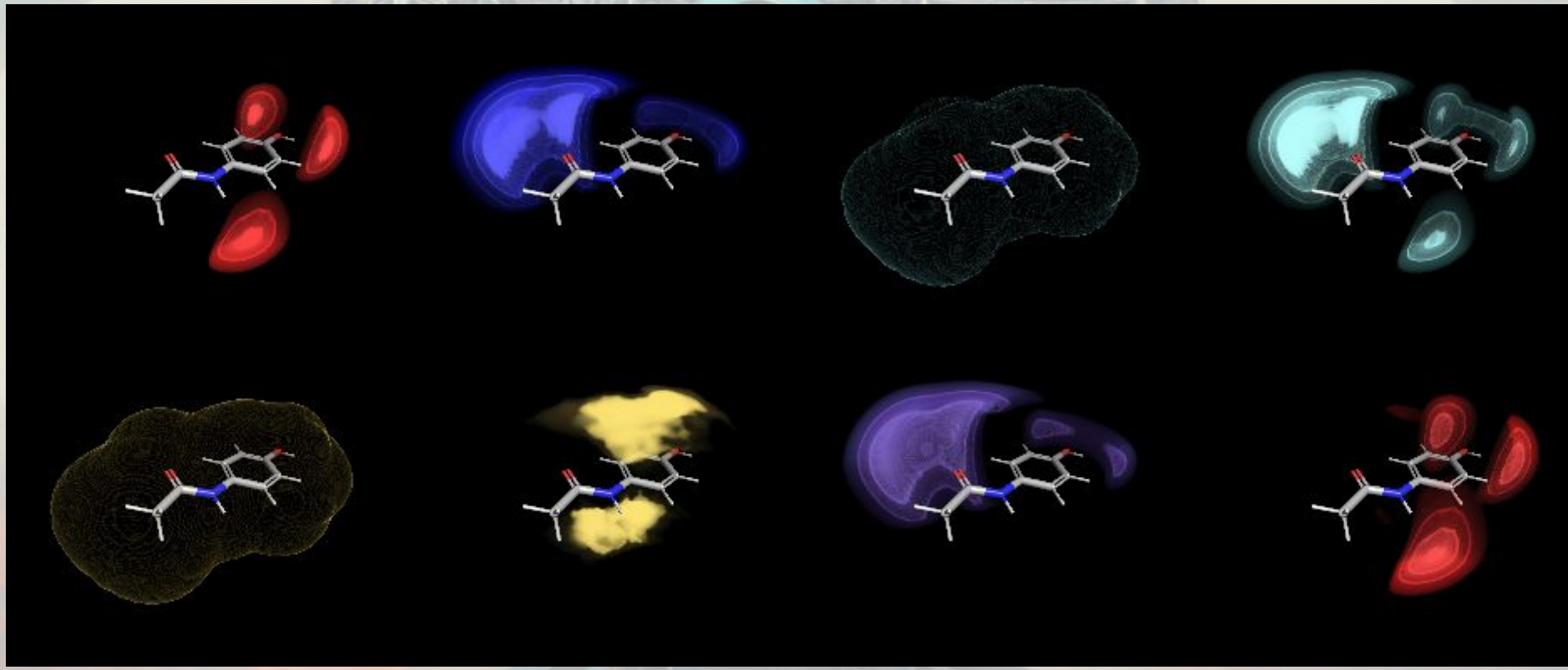


Initial Summary

- 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

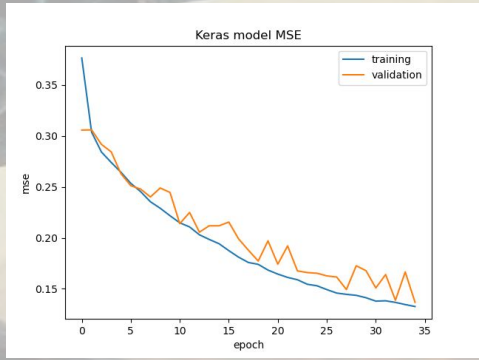
From 2D images with 3 channels → 3D images with 8 channels using GRID



DeepGRID Hyperparameters optimization

- A grid search has been used to test various combinations of Hyperparameters , including:
  - CNN filters, kernel sizes, number of dense layers, units per layer

- The model was run for 35 epochs and potentially could be run for longer for additional slight improvements



Removing CHEBI338620 as an outlier

- CHEBI338620 has an reported experimental lgBB 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

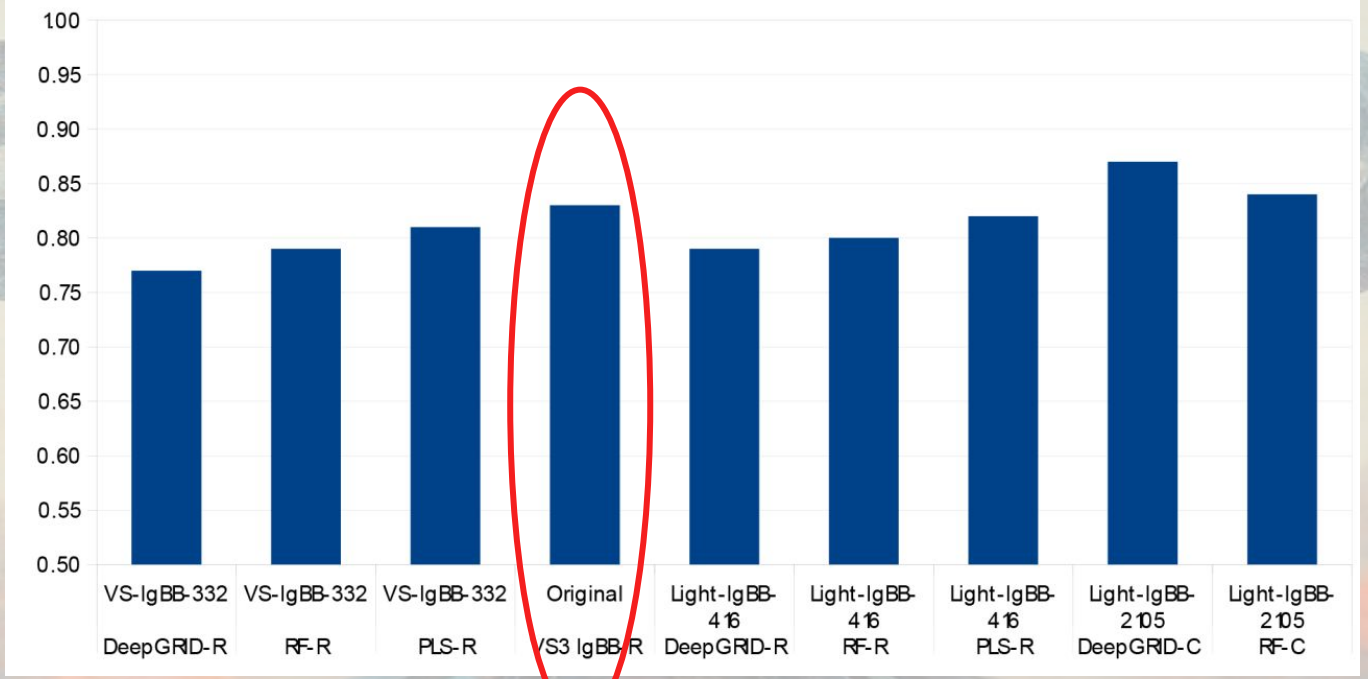
Light-IgBB-416 dataset is more diverse

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

VolSurf IgBB PLS model does a good job

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



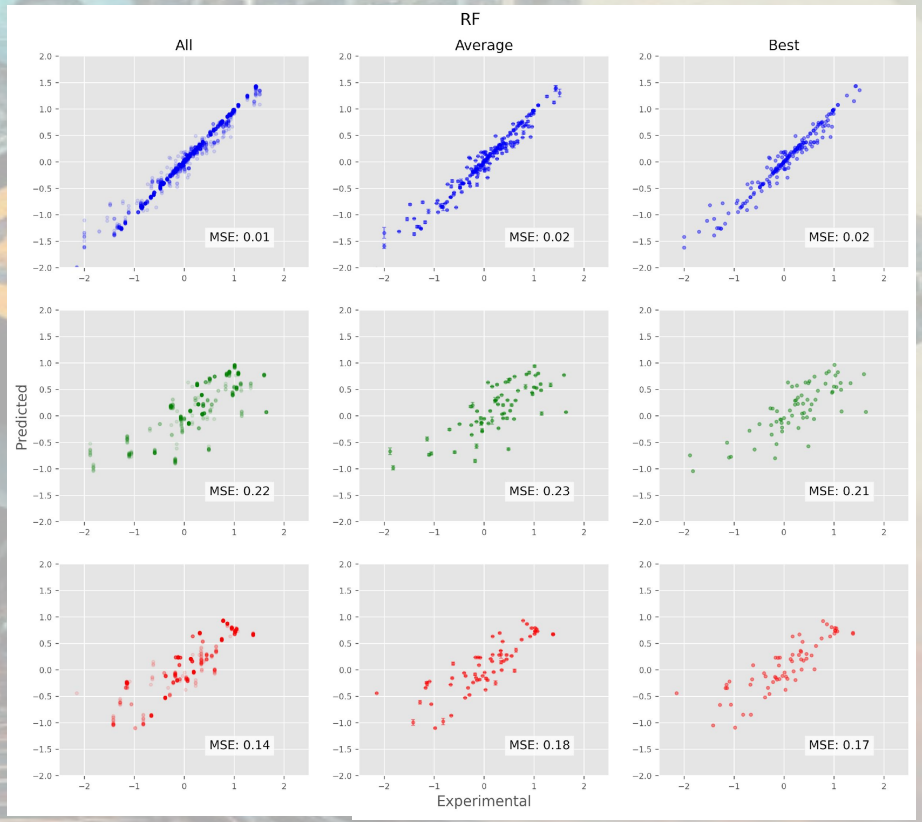
RF performance, VS-IgBB-332 dataset

The plots show Exp vs Pred for:

- All conformers
- Avg prediction across confs
- Best prediction by conf

The Avg prediction for the test set gives:

- MSE: 0.18



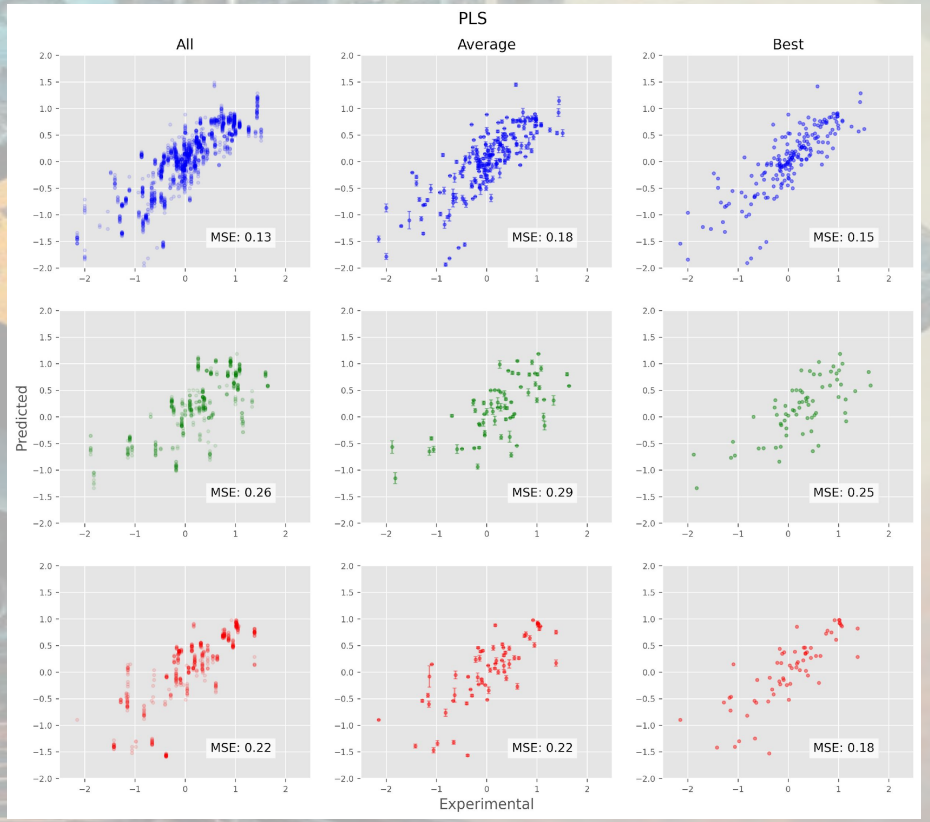
PLS performance, VS-IgBB-332 dataset

The plots show Exp vs Pred for:

- All conformers
- Avg prediction across confs
- Best prediction by conf

The Avg prediction for the test set gives:

- MSE: 0.22




GRID MIF

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} that is the product of three terms terms:

- E_r based on the distance between the target and the probe given in kcal/mol
- The other two, both ranging in the interval 0–1. They are dimensionless functions of the angles t and p made by the hydrogen bond (HB) at the target and the probe atoms respectively

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

Emin → dEmin

Er assumes relative values in case of interaction with a HB acceptor or donor complementary probe and is parametrized by two values: **Emin is the strongest hydrogen-bond attraction energy at the optimum position (Emin),** and half of the straight-line distance between donor and acceptor atom pairs which corresponds to the strongest hydrogen-bond attraction energy (Rmin).

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

The dataset is made of 66463 drug-like molecules

- We used GAMESS-US B3LYP/SVP (necessity of having a versatile basis set and method) to compute the Electrostatic Potential (EP) for each atom
- EP is converted to the so called dEmin value using linear equation derived so that for each AT all the resulting dEmin values always fall within an acceptable range

$$dEmin_{BH} = m_{BH} * EP + q_{BH}.$$

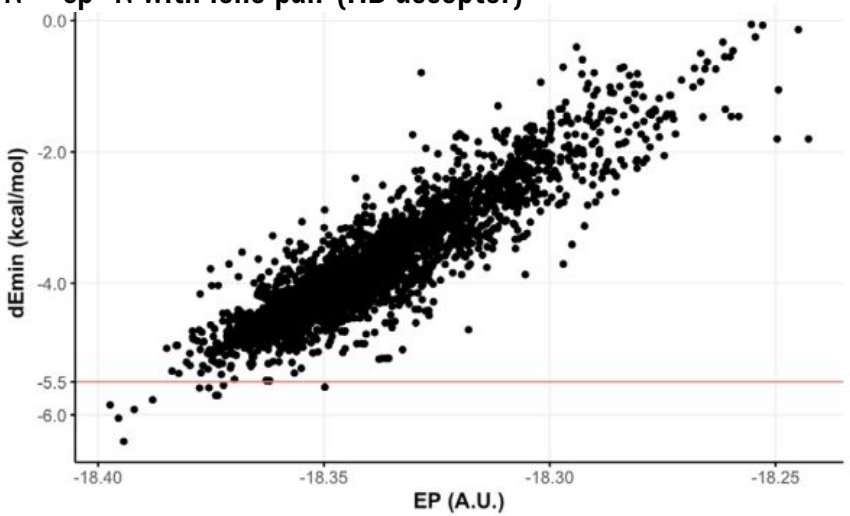
$$dEmin_{AH} = -m_{AH} * EP - q_{AH}.$$

22 equations, each one for each AtomType

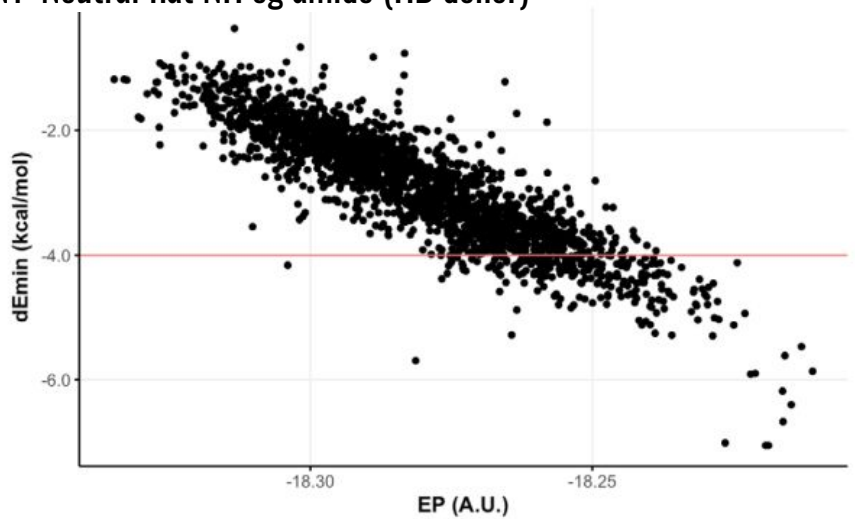
The dEmin is our label

Machine Learning and the GRID Force-Fields

N:= sp^2 N with lone pair (HB acceptor)



N1 Neutral flat NH eg amide (HB donor)



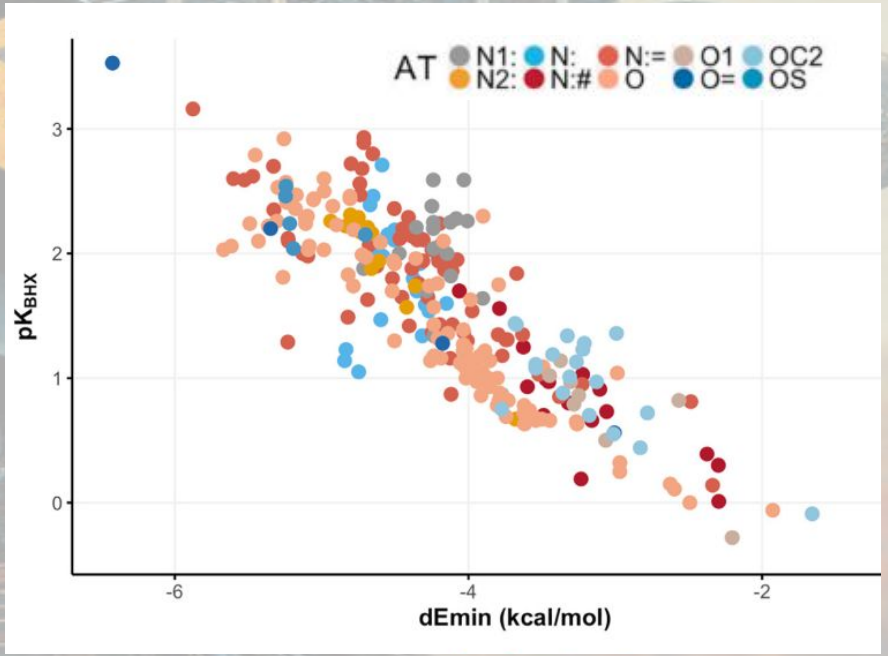
The red lines represent values of the traditional, static Emin of the GRID force field, namely -5.5 for N:= and -4.0 for N1 atom types. dEmin, dynamic Emin

Machine Learning and the GRID Force-Fields

Does chemically sound to use the dEmin in the the E_{HB} term ?

We decided to test the correlation of the proposed dEmin to those experimental hydrogen-bonding (HB) properties.

dEmin versus H-bond basicity scale for the Kenny dataset (279 atoms, R – Pearson = 0.85).

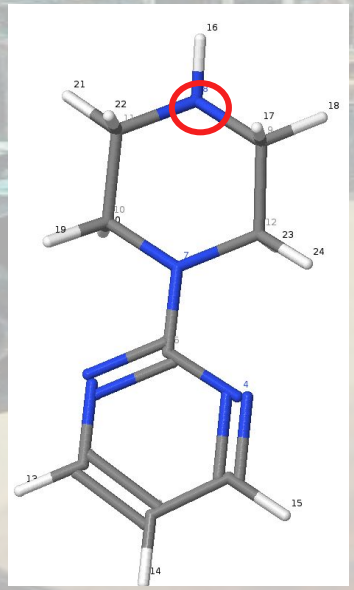


Machine Learning and the GRID Force-Fields

We have a good label, now we need to select the feature (descriptor) to use in the model

The molecular environment is described by a tree-structured molecular fingerprint with a length of 10 bond distances

```
0   1    8 N_3H 122
1   2    9 C.3 326   11 C.3 326
2   2   12 C.3 629   10 C.3 629
3   1    7 N.3_ar 1016
4   1    5 C.ar+ 1250
5   2    4 NPYM 1706    6 NPYM 1706
6   2    3 C.ar+ 1856    1 C.ar+ 1856
```

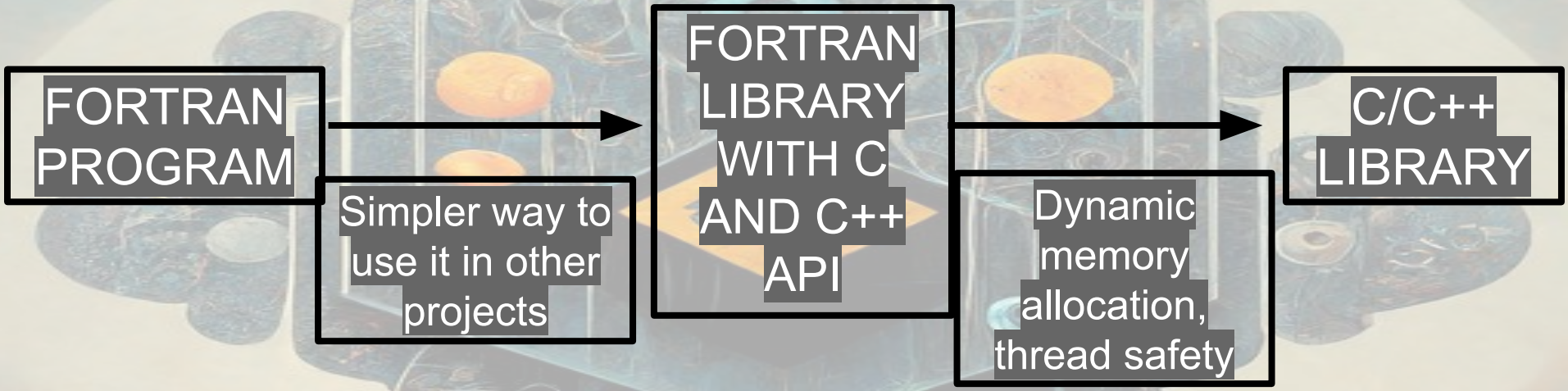


Machine Learning and the GRID Force-Fields

Using this approach, 22 PLS models were built relating atomic environment to dEmin for the HB GRID atom types (some of the models results are reported validated using leave-one-out crossvalidation)

AT	Description	H-bond type	Atoms	LV	R^2	Q^2	SDEC (kcal/Mol)	SDEP (kcal/Mol)
N:	sp3 (tertiary) nitrogen, accepting one H-bond	A	6954	9	0.92	0.88	0.56	0.41
N1:	sp3 (secondary) nitrogen, donating one hydrogen and accepting one H-bond	A	3941	8	0.91	0.84	0.24	0.49
		D	4776	7	0.96	0.92	0.30	0.53
N2:	sp3 (primary)nitrogen, donating up to two hydrogen and accepting one H-bond	A	3618	8	0.84	0.71	0.26	0.38
		D	4895	7	0.95	0.92	0.30	0.41
ON	oxygen of nitro or nitroso group, accepting up to two H-bond	A	4907	8	0.82	0.69	0.26	0.38
N:=	sp2 (aromatic) nitrogen, accepting one H-bond	A	27,140	12	0.91	0.89	0.35	0.47
N::	sp2 nitrogen with two lone pairs and one double bond	A	472	4	0.89	0.59	0.23	0.12
N:#	sp nitrogen	A	15,798	10	0.72	0.66	0.29	0.32

Machine Learning and the GRID Force-Fields

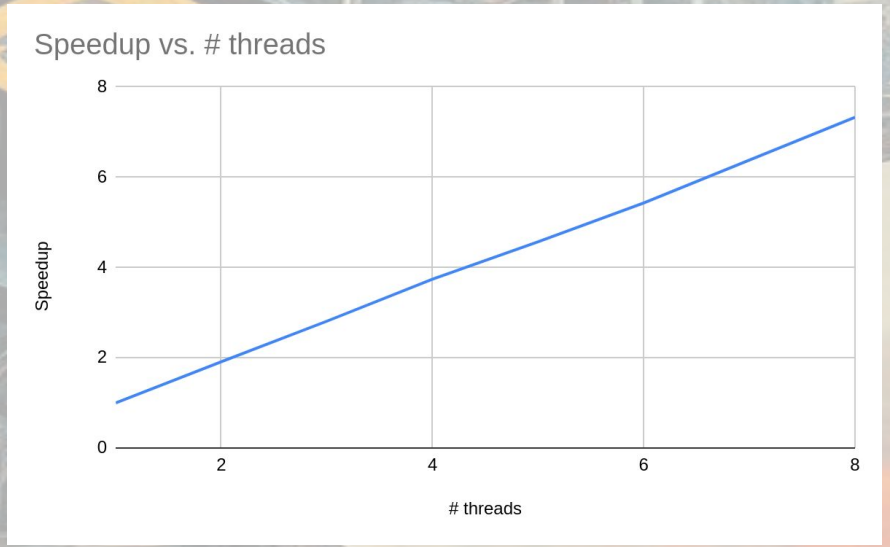


- pKa prediction for small molecules (and proteins)
- VolSurf, Almond and SHOP use molecular descriptors from 3D Molecular Interaction Fields (MIFs) produced by GRID
- MetaSite is a computational procedure that predicts metabolic transformations related to cytochrome-mediated reactions in phase I metabolism

Machine Learning and the GRID Force-Fields

- Easiest way is one thread for each probes (excellent speedup)
- One thread for each XY plane of the grid (big molecules)

# threads	Speedup
1	1
2	1.91
3	2.81
4	3.74
5	4.57
6	5.43
7	6.38
8	7.33



Intel Xeon E5420 2.50GHz


FORMULA SEARCH

A Formula search

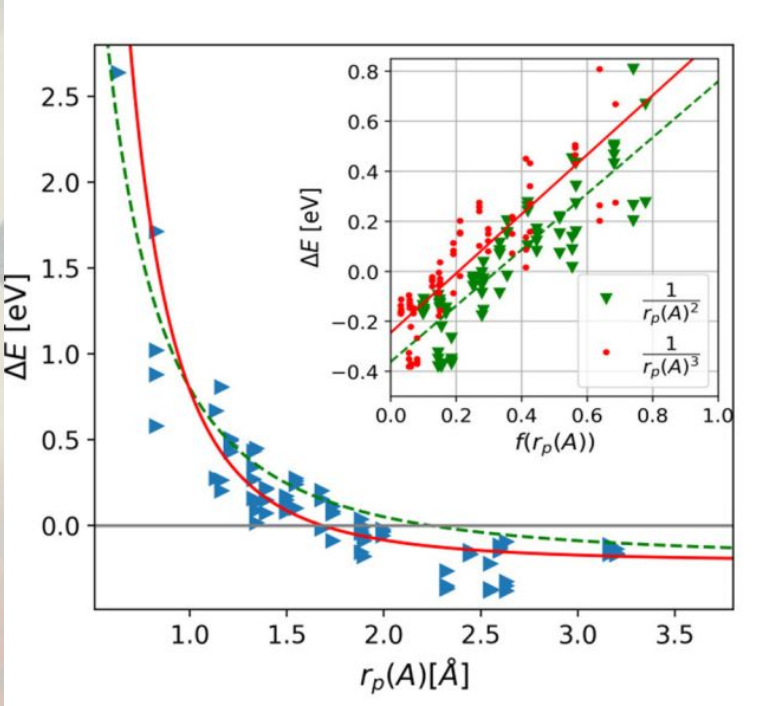
Formula search for crystal structure stability based on atomic properties.
Uses basic atomic properties to construct material features.
Employs machine learning methodology to construct formulas.
The final result is a transparent, human-readable mathematical formula.

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



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.


SCORING FUNCTION

A Scoring Function

ligand	ΔE^{FMO}	F2LE	E^{INT}	FE
1	−37.6	−1.6	−173.2	−7.2
2	−53.7	−2.1	−186.2	−7.2
3	−37.4	−1.5	−175.5	−7.0
4	−42.7	−1.7	−173.6	−6.9
5	−61.1	−2.5	−181.1	−7.5
6	−36.7	−1.5	−163.2	−6.8
7	−67.6	−3.1	−180.1	−8.2
8	−70.5	−3.2	−179.3	−8.2
9	−38.6	−1.8	−163.8	−7.4

△EFMO , F2LE, EINT and FE values computed for LR complexes formed by ligands 1–9 and hCA II. All energy values are in kcal/mol.

A Scoring Function

Ligand	HIE *	HIE-E *	logP
1	−38.9	−1.6	0.92
2	−37.9	−1.5	−0.01
3	−28.1	−1.1	−0.36
4	−30.6	−1.2	0.41
5	−35.0	−1.5	−0.28
6	−24.3	−1.0	0.68
7	−32.0	−1.5	0.6
8	−30.2	−1.4	0.32
9	−34.3	−1.6	0.46

* values in kcal/mol

Computed values for HIE, HIE-E (HIE/number of heavy atoms) and logP.