Modern Approaches in Classification of Covalent Organic Frameworks by Textural Properties Using JSL

Michael Nazarkovsky[1] and Felipe Lopes de Oliveira[2]

[1]Chemistry Department, Pontifical Catholic University of Rio de Janeiro (DQ-PUC Rio), Rio de Janeiro, Brazil

[2]Institute of Chemistry, Federal University of Rio de Janeiro (IQ-UFRJ), Rio de Janeiro, Brazil

Introduction

Covalent organic frameworks (COFs) are an emerging class of materials composed only of light elements such as carbon, nitrogen, oxygen, hydrogen, and boron.[1] COFs are designed in a bottom-up approach by the covalent bonding of one or more building blocks, following a well-defined geometric pattern dictated by the topology of the network resulting in organic, crystalline and nanoporous reticular material[2] (Fig.1).

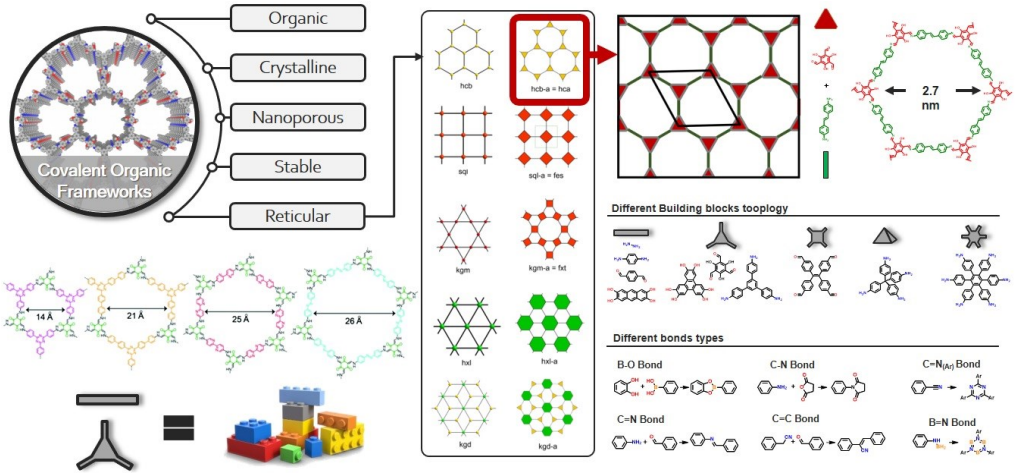


Fig.1. The sketch over the COFs structures: geometry, building blocks and the bonds types

There are various molecular geometries that can be used as building blocks and several types of chemical bonds that can be used to connect them.[3] In this way, it is possible to form structures with extremely specific characteristics, such as pore size that can be controlled in the order of angstroms and the physicochemical characteristics of the pore interior, only adjusting the building blocks used. Therefore, it is possible to think in the building blocks as Lego pieces, where different pieces can be used to build several structures with unique characteristics (Fig.2).

Due to this incredible modulation capacity, COFs can be used for several extremely interesting applications such as heterogeneous catalysis,[4–6] energy production, and storage, organic semiconductors,[7] chemical sensing,[8] thermal insulators,[9] and gas capture and storage,[10,11] among others[12] (Fig.3).

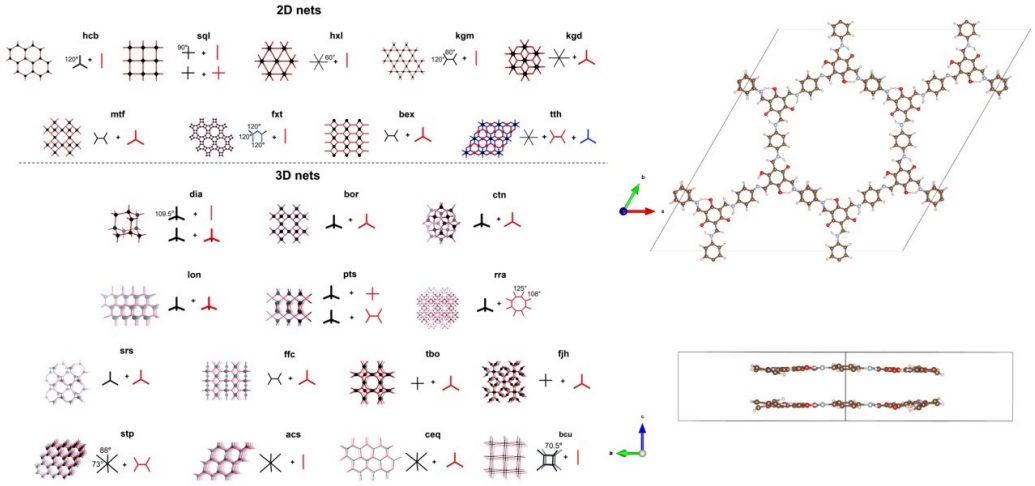


Fig.2. The concept of the building blocks for 2D and 3D structures

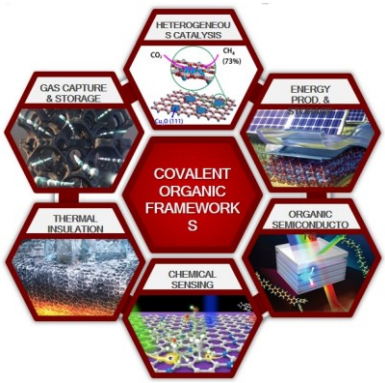


Fig.3. The variety of the COFs application

However, this large chemical diversity can also be a problem. The development of these materials usually follows a combination of chemical intuition and the use of available molecules. With several steps of synthesis and characterization that are repeated until a material with high crystallinity is obtained. This makes its development very time-consuming, expensive, and requires several trained professionals. And often, even with all this effort, it is not possible to obtain a material with the desired characteristics.

To overcome this complex experimental development, we propose a simple and fast computational approach based on the textural properties such as specific surface area, pore size, pore volume, and dimensionality to study these materials in the aim to accelerate their development.

In the presentation reported for the Summit, the structures were classified by means of unsupervised (Hierarchical Clustering, PCA) and supervised (Multiple Logistic Regression, Naive Bayes, KNN, SVM) methods coded in JSL (JMP Pro 15). The COFs were separated in two main groups, named 2D and 3D, based on their geometrical characteristics. The largest clusters were selected to perform supervised machine

learning stratifying the data by the structure (62.9% 2D and 37.1% 3D) to avoid disproportion in training/validation ratio (80%/20%). The 2D/3D classification by textural properties was successfully accomplished with the 100% accuracy (validation) for all models. Other metrics, such as Entropy R2, Generalized R2, Mean -log(p) were compared and discussed to select the optimal model. It was shown in the presentation to the present Summit 2021.

The paper basically focuses on the treatment of the whole dataset – it means that the data will pass the path from the cleaning until tuning the models. The presentation rested in the extraction of the clusters utilizing the typical machine learning for 100% Accuracy. Herewith, we claim the role of imbalanced output and how to resolve this issue dealing with all the data. Also, in the paper we used a deep learning algorithm, implementing a boosted Neural Network.

Methodology

For this work 590 COF structures with diverse building blocks, topologies, and dimensionalities were selected from a database of known as CURATED COFs,[13] which compiles several synthesized structures reported in the literature. The textural properties were calculated using the Zeo++[14] software, which performs geometric analysis of porous and crystalline materials based on the formation of the Voronoi networks and determines a set of structural descriptors for the selected COFs. The set of descriptors selected can be divides in three classes: i) pore size, ii) specific area and iii) channels number (Fig.4).

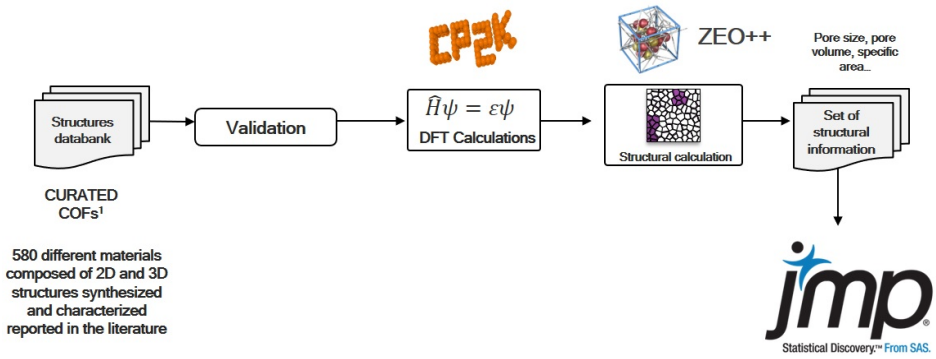


Fig.4. The computational pipeline for the COFs classification by machine learning

The pore size related descriptors were the large included sphere (LIS), which gives us information on the pore diameter of the material, the Large Free Sphere (LFS) and the Large free sphere path (LSFP). These last two are generally similar to the pore diameter, but for structures that have functional groups within the pore it can be

considerably different, so it is interesting to evaluate these three descriptors. Also related to the pore size the pore volume (V_p) in cm³/g is used to classify the structures.
The specific area related descriptors were the specific gravimetric area (SSA) and specific volumetric area (SSAV), in square meters per gram and in square meters per volume unit respectively.

The channel related descriptors were the number of accessible channels, the dimensionality of the channels, the size of the channels, and the accessible volume fraction (AVF), which is the percentage of the material composed by pores.

All the data are listed and defined:

COF_name – Name of the structure

LIS - Large included sphere (Pore Diameter) [Å]

LFS - Large free sphere [Å]

LSFP - large sphere free path [Å]

SSA - Specific surface area [m^2/g]

SSAV - Specific volumetric surface area m²/cm³

AVF - Accessible volume fraction

V_p - Pore volume cm³/g

N(chan) - Number of accessible channels

DimChan - Dimensionality of the channels (1D, 2D ou 3D)

ChanSize - Channel size [Å]

Results

The dataset was screened for missing, negative or zero values, whereas the set shrank to 580 structures. Thereupon, the outliers two outliers by criterion of quantile range were excluded from SSA (2 outlier) and Vp (2 outlier) belonging to the same two points at Q = 3 and the tail quantile of 0.1. Hence, the final number of the structures has become 576 with no other outliers (Table 1). Now, the cleaned dataset can be subjected to the machine learning experiments.

The scatter plots demonstrate strong collinearity within two sets of the inputs: between LIS, LFS and LSFP and between ChanSize X, Y and Z, which is critical while linear modeling. SSAV has a weak negative correlation with other parameters in 2D and it is a bit stronger for 3D keeping its negative character. Generally, the 2D dataset is characterized by more pronounced correlations, than 3D is.

Table 1. The final screening analysis after excluding outliers for 2D (A) and 3D (B) structures

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers Outliers (Count)	
LIS, Å	9.50553	29.506	-50.496	89.5074	0	
LFS, Å	8.60495	29.106	-52.898	90.6091	0	
LSFP, Å	9.48433	29.506	-50.581	89.571	0	
SSA, m2/g	1607.67	2767.96	-1873.2	6248.83	0	
SSAV, m2/cm3	985.999	1556.34	-725.01	3267.35	0	
AVF	0.25414	0.6182	-0.838	1.71037	0	
Vp, cm3/g	0.31535	1.56549	-3.4351	5.31593	0	
ChanSize X, Å	8.80503	29.1802	-52.32	90.3055	0	
ChanSize Y, Å	7.88985	28.91	-55.171	91.9706	0	
ChanSize Z, Å	8.80503	29.1802	-52.32	90.3055	0	(A)

Column	10% Quantile	90% Quantile	Low Threshold	High Threshold	Number of Outliers Outliers (Count)	
LIS, Å	5.78251	26.7402	-57.09	89.6132	0	
LFS, Å	4.33139	21.1113	-46.008	71.4509	0	
LSFP, Å	5.66509	26.2119	-55.975	87.8524	0	
SSA, m2/g	1036.31	7863.62	-19446	28345.6	0	
SSAV, m2/cm3	835.434	2339.99	-3678.2	6853.67	0	
AVF	0.10036	0.81916	-2.0561	2.97559	0	
Vp, cm3/g	0.09912	5.66305	-16.593	22.3549	0	
ChanSize X, Å	5.06368	26.7402	-59.966	91.7698	0	
ChanSize Y, Å	3.69199	21.1113	-48.566	73.3692	0	
ChanSize Z, Å	5.03536	26.6137	-59.7	91.3488	0	(B)

Table. 2. The scatterplot matrix with correlations for over the input parameters for 2D

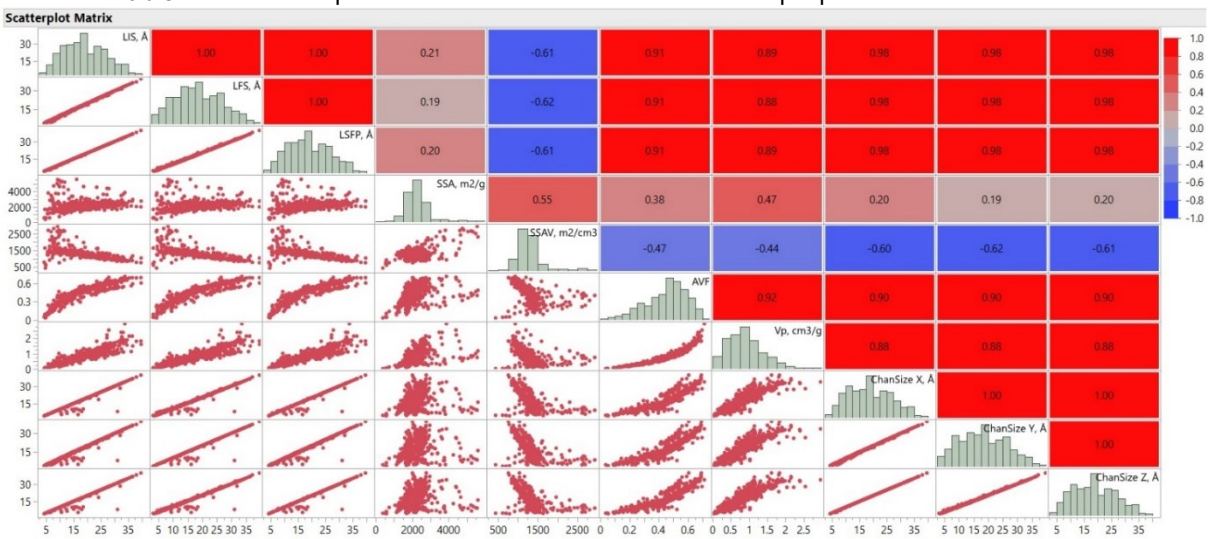
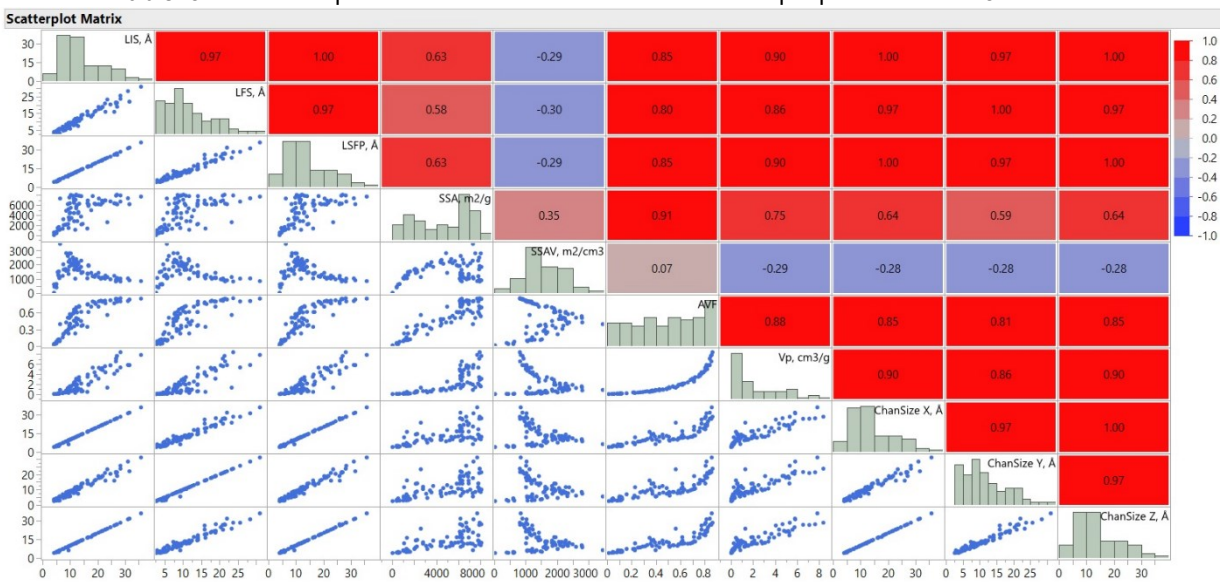


Table. 3. The scatterplot matrix with correlations for over the input parameters for 3D



The distribution of the structures to be classified has shown an unbalanced profile, where the 85%-portion is attributed to 2D. Thus, in order to build a correct predictive modeling, the training-validation split must be performed under stratification by “type”, as a binary response of 2D/3D (Fig.5, 6). Also, the profit table for “type” was adjusted (3D – 0.15 and 2D – 0.85).

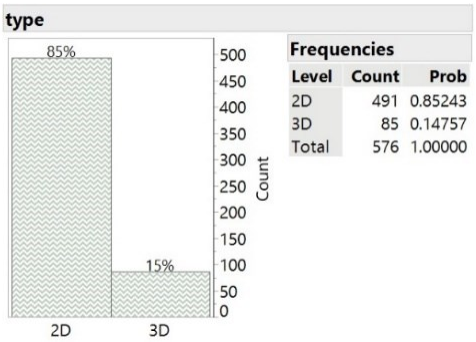


Fig.5. 2D/3D distribution histogram

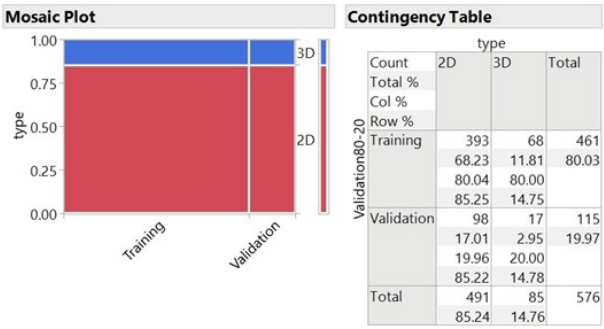


Fig.6. The training-validation split stratified by "type"

It is worth noting that non-continuous parameters, such a ChanDim and N(chan) can play an essential role as variables, having some issues in the regression (linear or logistic) with well-known instability during the parameters estimation. In other cases, for non-linear or non-parametric models, this feature is not encountered (Fig.7). It is seen that absolute majority of monodimensional channels (91.0%) is typical for 2D, whereas the 3D class possesses tridimensional and two-dimensional at the ratio of > 2.3.

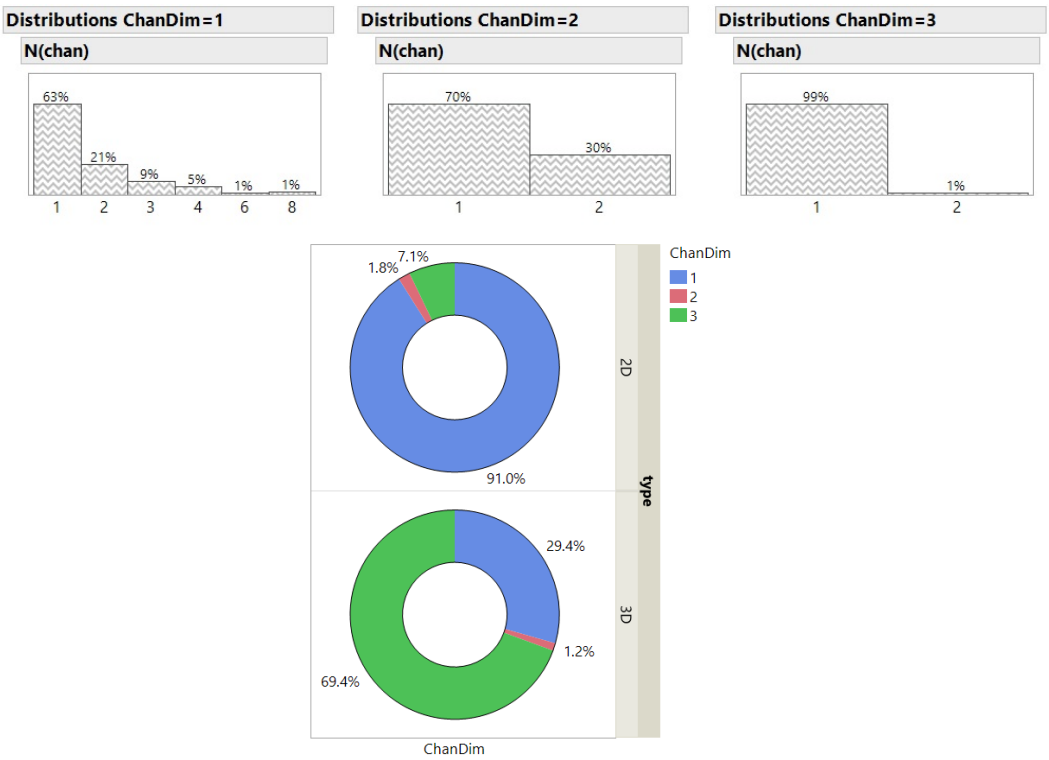


Fig.7. The distribution of the number of channels over the channels dimensionalities for 2D and 3D structures.

Starting with KNN (K_{max} = 100, Euclidean distance), the minimal misclassification rate or (1-Accuracy)× 0.01 of 8.7% with F1 = 0.995 are the results at K = 3 (Fig.8). The mean values of the neighbors' distances and the modes for 3D are shifted to higher values both for training and validation. To develop this model, all the variables (continuous and nominal) were involved. This leads us to conclusion that both types of

the structures can be considered as mean clusters, whose coordinates are the principal components.

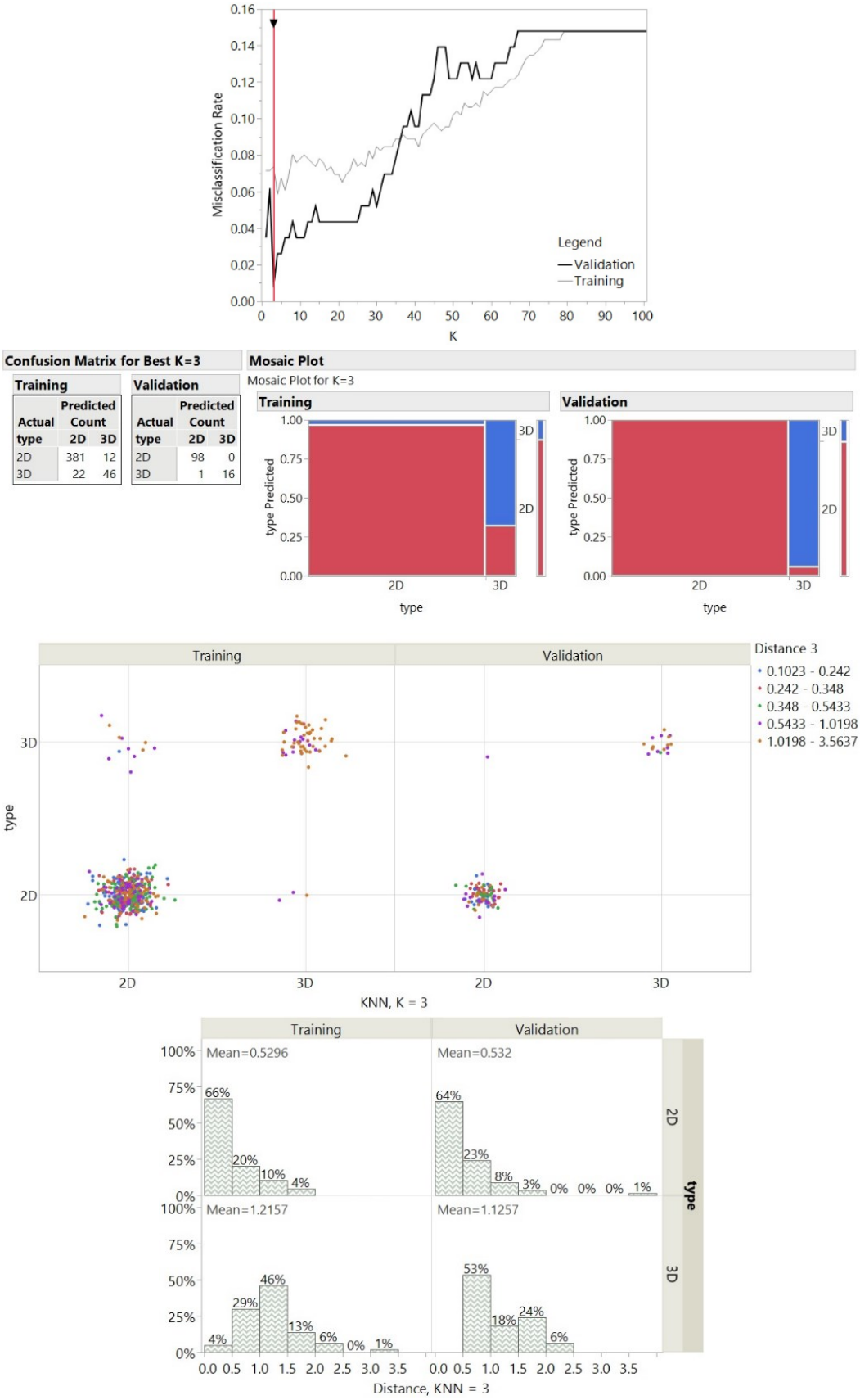


Fig.8. The summary of KNN: the model selection, the confusion matrices of training and validation, the distributions of the neighbors distances over the training-validation split

As for the multiple logistic regression, after removing all the variables at p > 0.05 and excluding those with unstable status in the parameters estimation, the model has got three main factors: LIS, SSA and SSAV (Fig.9). In other words, specific surface area and pore diameter are the main characteristics in the linear classification with a non-significant lack of fit (p = 1.000). The profit matrix has given 7.83% of balanced misclassification, whereas F1 = 0.952. In the contrast to KNN, this model has given less accuracy, but it involved only three parameters to distinguish 2D/3D COFs. In the other hand, the accuracy for KNN during validation has turned out to be worse than at training - as opposed to the case of multiple logistic regression, where the training has higher misclassification (8.89%) than validation.

Source	LogWorth	PValue
SSA, m2/g	43.176	0.00000
LIS, Å	26.957	0.00000
SSAV, m2/cm3	16.104	0.00000

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	456	76.315587	152.6312
Saturated	459	0.000000	Prob>ChiSq
Fitted	3	76.315587	1.0000

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-5.2346197	1.1716133	19.96	<.0001*
LIS, Å	0.45331148	0.0620989	53.29	<.0001*
SSA, m2/g	-0.0023846	0.000362	43.40	<.0001*
SSAV, m2/cm3	0.00499744	0.0009125	30.00	<.0001*

For log odds of 2D/3D

Decision Matrix

Training

Actual type	Decision Count 2D	Decision Count 3D
2D	359	34
3D	7	61

Actual type	Decision Rate 2D	Decision Rate 3D
2D	0.913	0.087
3D	0.103	0.897

Misclassification Rate
0.0889

Validation

Actual type	Decision Count 2D	Decision Count 3D
2D	90	8
3D	1	16

Actual type	Decision Rate 2D	Decision Rate 3D
2D	0.918	0.082
3D	0.059	0.941

Misclassification Rate
0.0783

Specified Profit Matrix

Actual	Decision 2D	Decision 3D
2D	0	-0.176
3D	-1	0

Confusion Matrix

Training

Actual type	Predicted Count 2D	Predicted Count 3D
2D	387	6
3D	25	43

Validation

Actual type	Predicted Count 2D	Predicted Count 3D
2D	97	1
3D	6	11

Fig.9. The summary on the Multiple Logistic Regression

The Quadratic Discriminant Analysis involving only three main variables from Multiple Logistic Regression (SSA, LIS and SSAV) basically demonstrates quite weak classification – the improved decision matrix on validation gives MR 13.91% at F1 = 0.912 (Fig.10).

Decision Tree after 14 splits showed that ChanSize of all three directions (X, Y and Z) and LSFP did not contribute to the model (Fig. 11) and, hence, the model was re-launched and tuned excluding these parameters (Fig. 12). Simplified model has not lost the accuracy (F1 = 0.979) and the profit value in the respective matrix – its validation misclassification rate is 3.48% with balanced distribution of the misclassified data, and less than for training (4.34%) and F1 = 0.974. The model has shown to be more accurate than multiple logistic regression and does not overfit the data.

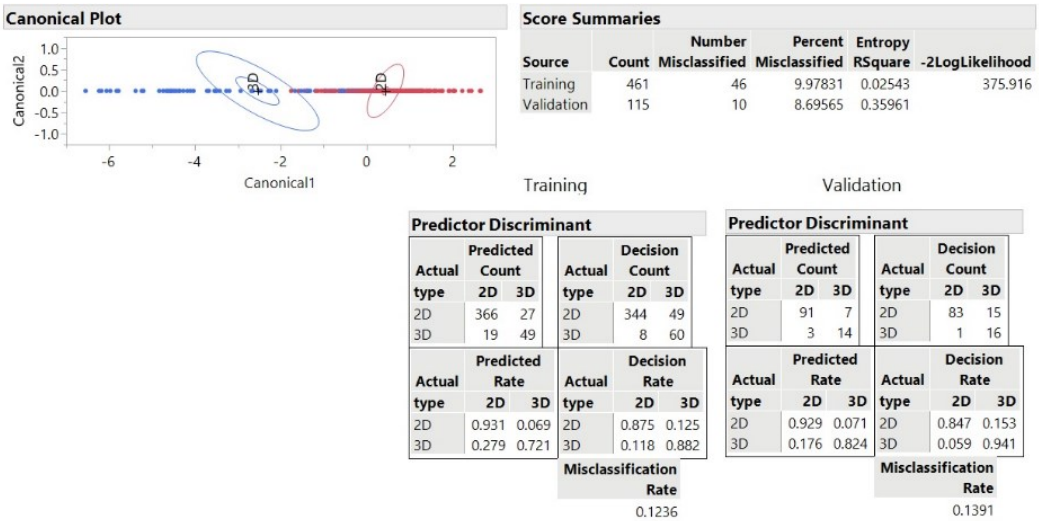


Fig.10. The Discriminant Analysis summary

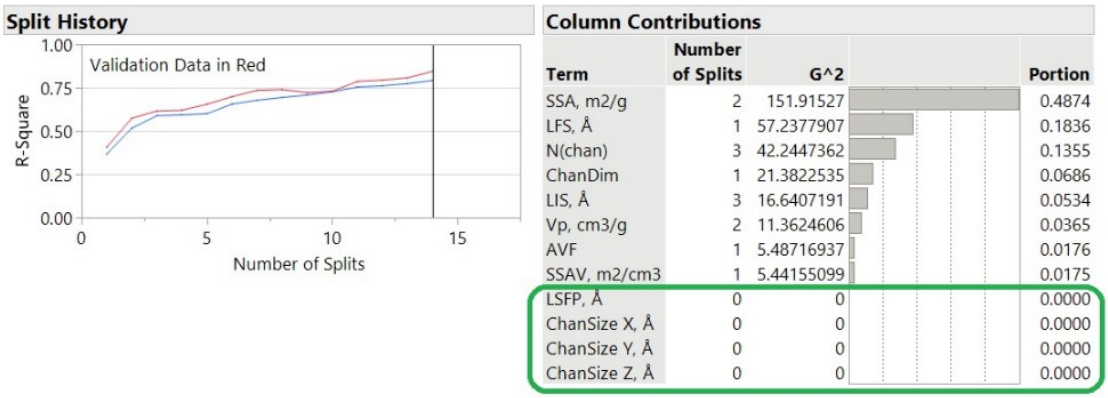


Fig.11. The overall model of Decision Tree to classify 2D/3D structures

Similar situation is encountered with Bootstrap Forest (10 terms sampled per split, 12 terms, the minimum size split = 5) – two variables related to the ChanSize, namely X and Z, contribute poorly to the model (< 0.01), giving quite high MR (6.09%) and F1 (0.963). (Fig.13). Thus, the tuned Bootstrap Forest (5 terms per split, 10 terms, the minimal split = 5) revealed a decrease of the trees (only two trees), the misclassification rate in validation (3.48%) and F1 = 0.979 (Fig.14) due to the re-distribution in 2D classification: from 6.1% of misclassified after tuning the MR was reduced down to 3.1%.

The Boosted Tree (140 layers, 5 splits per tree, the learning rate = 0.1) let us to exclude two last variables (ChanSize Z and LSFP) for more adequate fitting (Fig.15) and the resulted tuned (adjusted) Boosted Tree gave more reduced number of the layers (88 layers, 10 splits per tree, learning rate = 0.1) (Fig.16, 17). Moreover, the metrics of the model underwent the significant improvement fur to better attribution of 2D structures:

MR for 2D in the initial full model was 3.1%, whereas after tuning, MR became 1.0%, which does not change after the 36th layer.

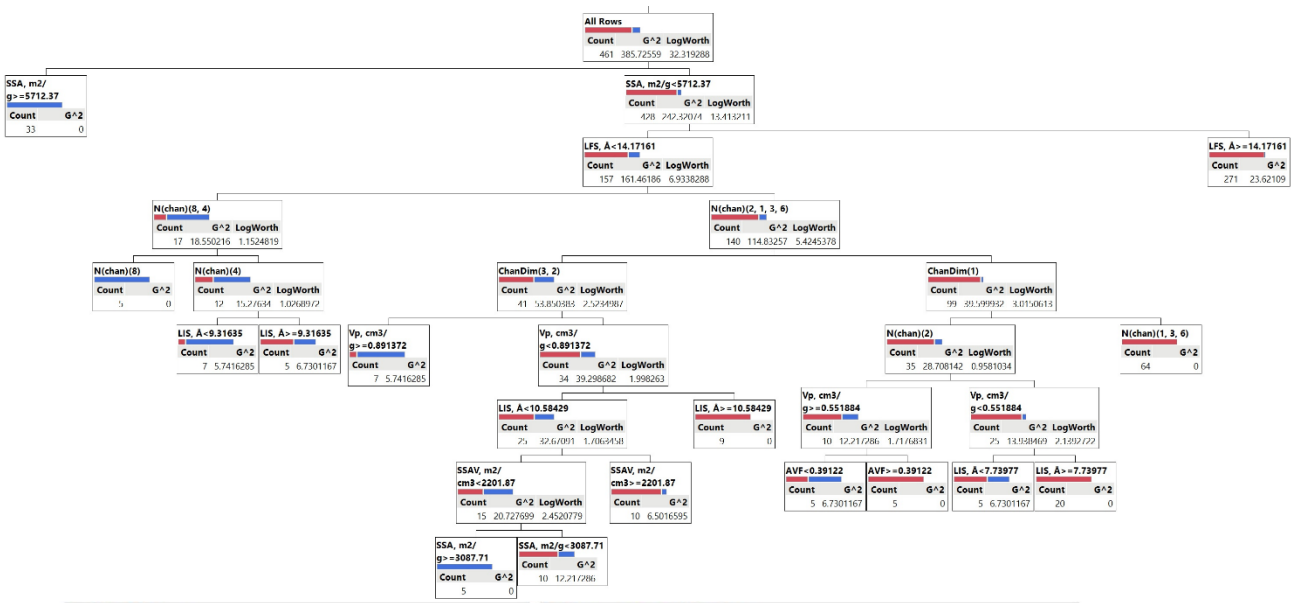
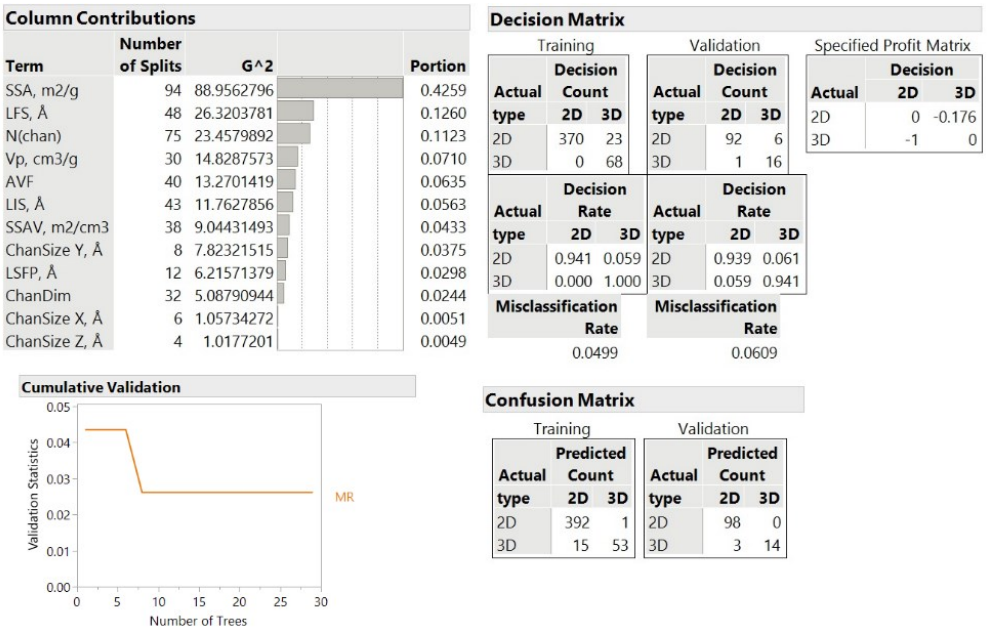


Fig.12. Tuned Decision Tree model's summary: the split tree structure, split history, the variable contribution, decision and confusions matrices



Column Contributions

Term	Number of Splits	G^2	Portion
SSA, m2/g	94	88.9562796	0.4259
LFS, Å	48	26.3203781	0.1260
N(chan)	75	23.4579892	0.1123
Vp, cm3/g	30	14.8287573	0.0710
AVF	40	13.2701419	0.0635
LIS, Å	43	11.7627856	0.0563
SSAV, m2/cm3	38	9.04431493	0.0433
ChanSize Y, Å	8	7.82321515	0.0375
LSFP, Å	12	6.21571379	0.0298
ChanDim	32	5.08790944	0.0244
ChanSize X, Å	6	1.05734272	0.0051
ChanSize Z, Å	4	1.0177201	0.0049

Decision Matrix

Training

Actual type	Decision Count 2D	3D
2D	370	23
3D	0	68

Actual type	Decision Rate 2D	3D
2D	0.941	0.059
3D	0.000	1.000

Misclassification Rate: 0.0499

Validation

Actual type	Decision Count 2D	3D
2D	92	6
3D	1	16

Actual type	Decision Rate 2D	3D
2D	0.939	0.061
3D	0.059	0.941

Misclassification Rate: 0.0609

Specified Profit Matrix

Actual	Decision 2D	3D
2D	0	-0.176
3D	-1	0

Confusion Matrix

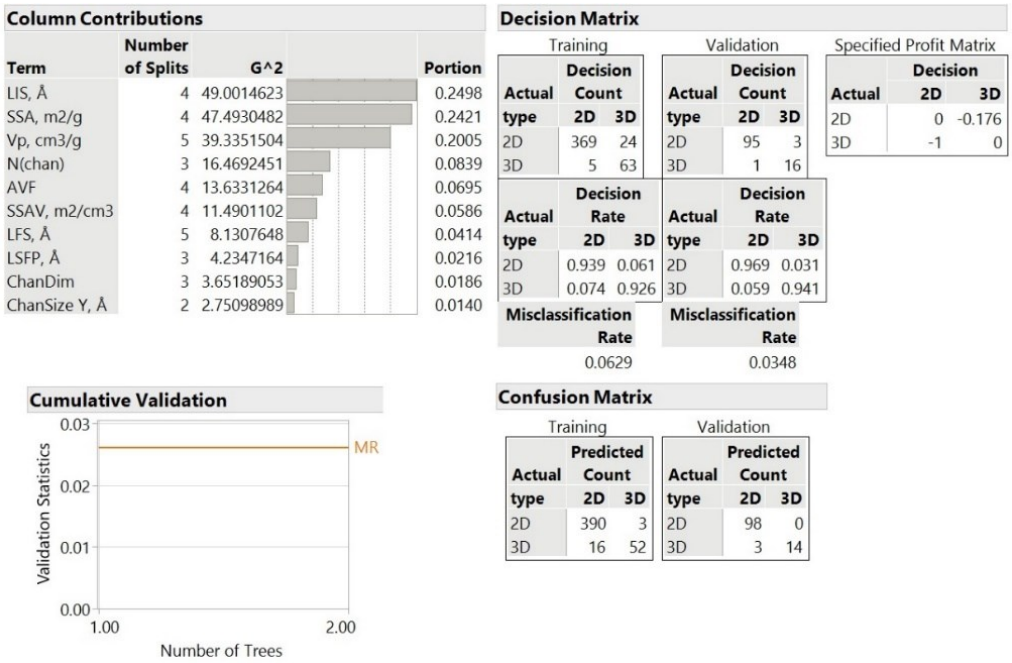
Training

Actual type	Predicted Count 2D	3D
2D	392	1
3D	15	53

Validation

Actual type	Predicted Count 2D	3D
2D	98	0
3D	3	14

Fig.13. The overall Bootstrap Forest model's summary: the variable contribution, cumulative validation progress, decision and confusions matrices.



Column Contributions

Term	Number of Splits	G^2	Portion
LIS, Å	4	49.0014623	0.2498
SSA, m2/g	4	47.4930482	0.2421
Vp, cm3/g	5	39.3351504	0.2005
N(chan)	3	16.4692451	0.0839
AVF	4	13.6331264	0.0695
SSAV, m2/cm3	4	11.4901102	0.0586
LFS, Å	5	8.1307648	0.0414
LSFP, Å	3	4.2347164	0.0216
ChanDim	3	3.65189053	0.0186
ChanSize Y, Å	2	2.75098989	0.0140

Decision Matrix

Training

Actual type	Decision Count 2D	3D
2D	369	24
3D	5	63

Actual type	Decision Rate 2D	3D
2D	0.939	0.061
3D	0.074	0.926

Misclassification Rate: 0.0629

Validation

Actual type	Decision Count 2D	3D
2D	95	3
3D	1	16

Actual type	Decision Rate 2D	3D
2D	0.969	0.031
3D	0.059	0.941

Misclassification Rate: 0.0348

Specified Profit Matrix

Actual	Decision 2D	3D
2D	0	-0.176
3D	-1	0

Confusion Matrix

Training

Actual type	Predicted Count 2D	3D
2D	390	3
3D	16	52

Validation

Actual type	Predicted Count 2D	3D
2D	98	0
3D	3	14

Fig.14. Tuned Bootstrap Forest model's summary: the variable contribution, cumulative validation progress, decision and confusions matrices.

Column Contributions

Term	Number of Splits	G^2	Portion
ChanDim	266	38566.3038	0.6160
N(chan)	182	12141.5987	0.1939
SSA, m2/g	49	2960.58252	0.0473
SSAV, m2/cm3	61	2613.02466	0.0417
AVF	46	2386.14396	0.0381
LFS, Å	34	1496.32315	0.0239
Vp, cm3/g	22	846.574883	0.0135
LIS, Å	19	626.325979	0.0100
ChanSize X, Å	9	537.024361	0.0086
ChanSize Y, Å	9	323.122722	0.0052
ChanSize Z, Å	2	83.4297327	0.0013
LSFP, Å	1	23.9562445	0.0004

Decision Matrix

Training

Actual type	Decision Count 2D	3D
2D	389	4
3D	1	67

Actual type	Decision Rate 2D	3D
2D	0.990	0.010
3D	0.015	0.985

Misclassification Rate 0.0108

Validation

Actual type	Decision Count 2D	3D
2D	95	3
3D	1	16

Actual type	Decision Rate 2D	3D
2D	0.969	0.031
3D	0.059	0.941

Misclassification Rate 0.0348

Specified Profit Matrix

Actual	Decision 2D	3D
2D	0	-0.176
3D	-1	0

Confusion Matrix

Training

Actual type	Predicted Count 2D	3D
2D	392	1
3D	4	64

Validation

Actual type	Predicted Count 2D	3D
2D	97	1
3D	3	14

Fig.15. The Boosted Tree summary for the full model: the variable contribution, decision and confusions matrices

Column Contributions

Term	Number of Splits	G^2	Portion
ChanDim	173	14788.2565	0.3976
N(chan)	306	10700.569	0.2877
SSA, m2/g	87	2708.25816	0.0728
AVF	73	2267.8157	0.0610
SSAV, m2/cm3	68	2101.46698	0.0565
LFS, Å	36	1497.87456	0.0403
Vp, cm3/g	58	1186.98528	0.0319
LIS, Å	45	980.941258	0.0264
ChanSize X, Å	15	499.499875	0.0134
ChanSize Y, Å	19	466.658052	0.0125

Decision Matrix

Training

Actual type	Decision Count 2D	3D
2D	392	1
3D	0	68

Actual type	Decision Rate 2D	3D
2D	0.997	0.003
3D	0.000	1.000

Misclassification Rate 0.0022

Validation

Actual type	Decision Count 2D	3D
2D	97	1
3D	1	16

Actual type	Decision Rate 2D	3D
2D	0.990	0.010
3D	0.059	0.941

Misclassification Rate 0.0174

Specified Profit Matrix

Actual	Decision 2D	3D
2D	0	-0.176
3D	-1	0

Confusion Matrix

Training

Actual type	Predicted Count 2D	3D
2D	392	1
3D	0	68

Validation

Actual type	Predicted Count 2D	3D
2D	97	1
3D	2	15

Fig.16. The Boosted Tree summary after tuning: the variable contribution, decision and confusions matrices

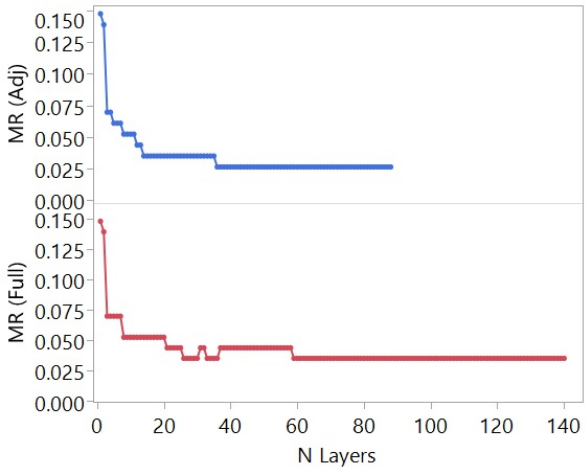
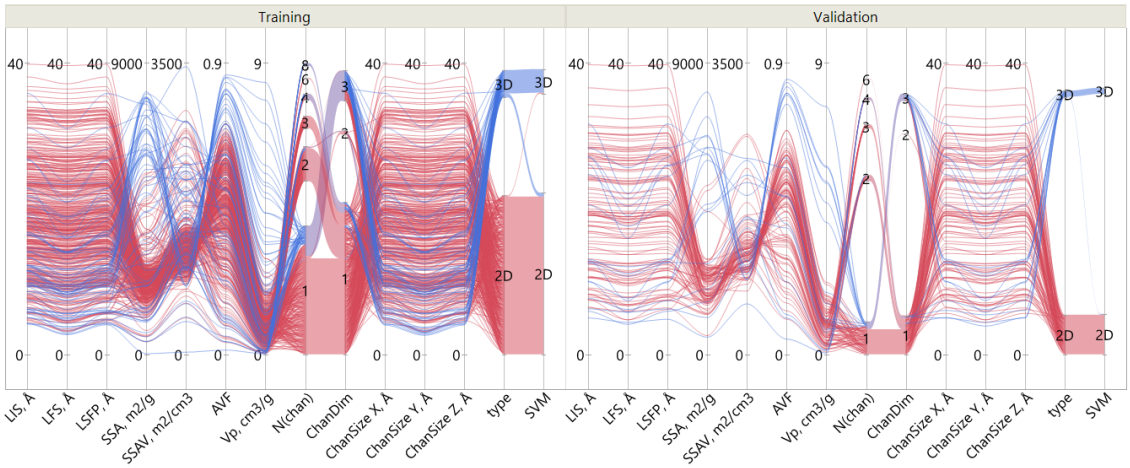


Fig.17. The cumulative MR progress with the layers for the full model – MR(Full) and for the tuned model – MR (Adj)

Regarding the Support Vector Machine which comprises all the variables, this model misclassified the structures at the same MR (<1%), as KNN did with quite balanced distribution in the confusion matrix at validation (Fig.18). But a huge difference between validation and training may speak for underfitting, although such a difference in the case of KNN is more pronounced. Thus, in spite of the equal F1 at validation (0.995), the model underfits less than KNN.



Training

Actual type	Predicted Rate 2D	Predicted Rate 3D
2D	0.990	0.010
3D	0.147	0.853

Misclassification Rate: 0.0304

Actual type	Predicted Count 2D	Predicted Count 3D
2D	389	4
3D	10	58

Validation

Actual type	Predicted Rate 2D	Predicted Rate 3D
2D	1.000	0.000
3D	0.059	0.941

Misclassification Rate: 0.0087

Actual type	Predicted Count 2D	Predicted Count 3D
2D	98	0
3D	1	16

Fig.18. The Support Vector Machine results: the parallel plots of the variables vs. real distribution and the predicted 2D/3D distribution - over training and validation. The confusion matrices for training and validation.

After arranging all the models by their metrics into Tables 4, we can conclude that during the training, the Tree family has shown itself more efficient by the key KPIs. Support Vector Machines stays on the third position in training, but is a leader model at the validation. Comparing the ROCs (Fig. 19), the main three models reveal their effectiveness: Boosted Tree, SVM and Bootstrap Forest.

Table 4. The model comparison by performance at training

Model	Entropy R^2	Generalized R^2	Mean -Log p	MR	Average Profit	AUC
Boosted Tree	0.9690	0.9799	0.013	0.0022	$-4 \cdot 10^{-4}$	0.9999
Classification Tree	0.7916	0.8545	0.0872	0.0304	-0.013	0.9828
Support Vector Machines	0.7807	0.8461	0.0917	0.0282	-0.015	0.9920
Bootstrap Forest	0.7391	0.8136	0.1091	0.0412	-0.02	0.9865
Logistic Regression	0.6043	0.7001	0.1655	0.0672	-0.028	0.9533
Discriminant Analysis	0.0254	0.0371	0.4077	0.0998	-0.036	0.9456

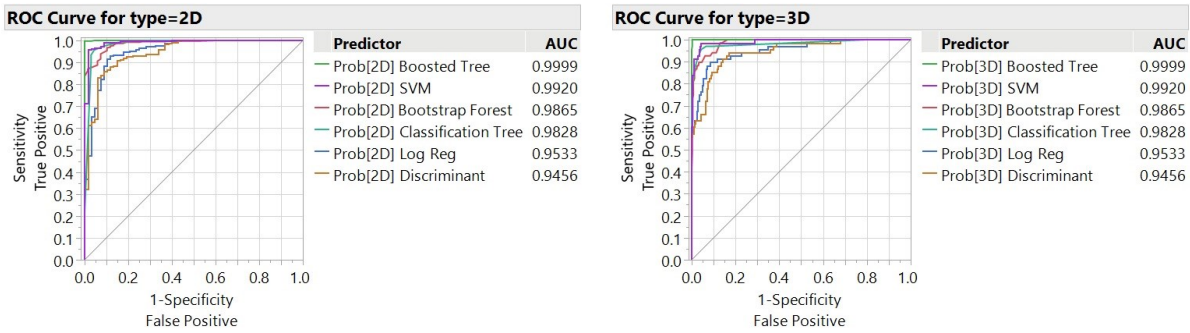


Fig.19. The ROC curves for training over all the involved models

The situation is a bit different at the validation: Classification Tree, SVM and Boosted Tree keep leadership as on ROCs (Fig.20), as by tabulated KPIs (Table 5). The only significant difference between the AUCs for SVM is demonstrated by the Discriminant Analysis, however, *p* is not small enough (0.0481) to reject H_o about the non-significant difference between the models by their AUC (Table 6).

Thus, for the sufficient classification of the COFs structures by the textural properties, we can recommend KNN (K = 3), SVM (γ = 0.5, cost = 1, number of the support vectors = 147), Classification Tree (14 splits).

Table 5. The model comparison by performance at validation

Model	Entropy R^2	Generalized R^2	Mean -Log p	MR	Average Profit	AUC	F1
Support Vector Machines	0.8469	0.8956	0.0641	0.0174	-0.01	0.9940	0.995
Classification Tree	0.8439	0.8934	0.0654	0.0174	-0.013	0.9964	0.979
Boosted Tree	0.8169	0.8735	0.0767	0.0261	-0.01	0.9934	0.990
Bootstrap Forest	0.8130	0.8707	0.0783	0.0261	-0.013	0.9928	0.979
Logistic Regression	0.6928	0.7762	0.1287	0.0609	-0.021	0.9838	0.952
Discriminant Analysis	0.3596	0.4585	0.2683	0.0870	-0.032	0.9706	0.912

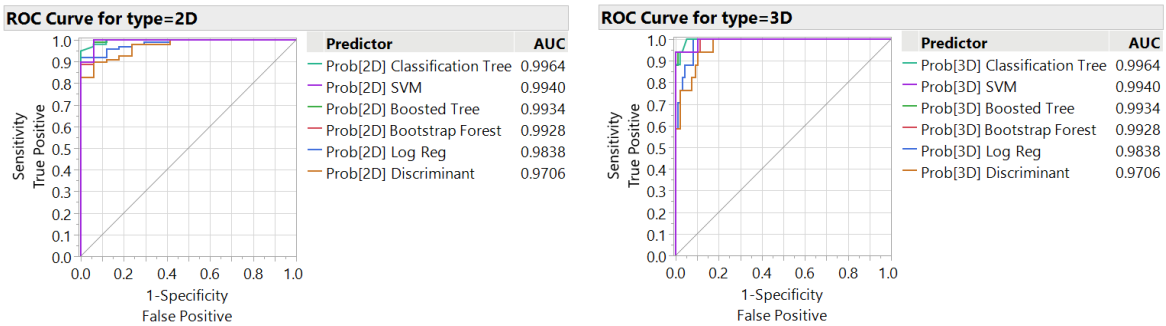


Fig.20. The ROC curves for validation over all the involved models

Table 6. The AUC difference analysis

Model	vs. Model	AUC Difference	Std Error	Lower 95%	Upper 95%	χ^2	$p>\chi^2$
Bootstrap Forest	SVM	-0.001	0.0096	-0.020	0.0177	0.0155	0.9009
Boosted Tree	SVM	-0.001	0.0023	-0.005	0.0038	0.0708	0.7902
Logistic Regression	SVM	-0.010	0.0084	-0.027	0.0063	1.4723	0.2250
Discriminant Analysis	SVM	-0.023	0.0118	-0.047	-0.000	3.9061	0.0481*
Classification Tree	SVM	0.0024	0.0072	-0.012	0.0164	0.1127	0.7371
Bootstrap Forest	Logistic Regression	0.0090	0.0112	-0.013	0.0310	0.6423	0.4229
Boosted Tree	Logistic Regression	0.0096	0.0081	-0.006	0.0254	1.4186	0.2336
Bootstrap Forest	Discriminant Analysis	0.0222	0.0159	-0.009	0.0534	1.9472	0.1629
Boosted Tree	Discriminant Analysis	0.0228	0.0115	0.0002	0.0454	3.9178	0.0478*
Logistic Regression	Discriminant Analysis	0.0132	0.0082	-0.003	0.0293	2.6001	0.1069
Bootstrap Forest	Classification Tree	-0.004	0.0050	-0.013	0.0062	0.5203	0.4707
Boosted Tree	Classification Tree	-0.003	0.0068	-0.016	0.0104	0.1925	0.6608
Logistic Regression	Classification Tree	-0.013	0.0094	-0.031	0.0059	1.7836	0.1817
Discriminant Analysis	Classification Tree	-0.026	0.0146	-0.055	0.0029	3.107	0.0780
Bootstrap Forest	Boosted Tree	-0.001	0.0092	-0.019	0.0174	0.0043	0.9480

Finally, all the available variables served for the boosted Neural Network of 39 sigmoidal activation functions incorporated in a single layer (Fig.21) after one tour of iterations with squared penalty method. Comparing training and validation matrices, one can easily see that neither over- nor underfitting occurred – this also is proven by minor difference between the Entropy R^2 and Generalized R^2. The respective interactive html-coded calculator (profiler) and the estimates for each activation function summary are uploaded to a Git-Hub repository: *https://github.com/Nazarkovsky/COFs-dimensionality-prediction.-Boosted-Neural-Network.*

Training

type

Measures	Value
Generalized RSquare	0.9963629
Entropy RSquare	0.9943244
Misclassification Rate	0

Confusion Matrix

Actual type	Predicted Count 2D	3D
2D	393	0
3D	0	68

Confusion Rates

Actual type	Predicted Rate 2D	3D
2D	1.000	0.000
3D	0.000	1.000

Validation

type

Measures	Value
Generalized RSquare	0.9974741
Entropy RSquare	0.996053
Misclassification Rate	0

Confusion Matrix

Actual type	Predicted Count 2D	3D
2D	98	0
3D	0	17

Confusion Rates

Actual type	Predicted Rate 2D	3D
2D	1.000	0.000
3D	0.000	1.000

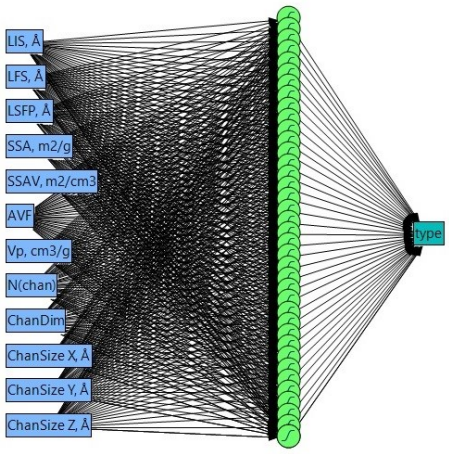


Fig.21. The Neural Network summary, confusion matrix and the diagram

Conclusion

As a conclusion, we recommend to test the most efficient models, SVM, KNN and Neural Network on new data to improve the performance of the proposed techniques, if necessary. After testing the models with thousands of the structures, we can take the final decision on the selection of the preferred machine learning algorithm (or an ensemble of the models) to be deployed and utilized as an application or interactive table to predict the dimensionality of the covalent organic frameworks. This study is helpful for the future predictive modeling of the physicochemical properties for each type of COFs.

References

[1] A.P. Cote, A.P. Côte, A.I. Benin, N.W. Ockwig, M. O'Keeffe, A.J. Matzger, O.M. Yaghi, *Science (80-.).* **2005**, *310*, 1166–1170.

[2] C.S. Diercks, O.M. Yaghi, The atom, the molecule, and the covalent organic framework., *American Association for the Advancement of Science, [***2017***,]*.

[3] X. Feng, X. Ding, D. Jiang, *Chem. Soc. Rev.* **2012**, *41*, 6010–6022.

[4] R.S.B. Gonçalves, A.B.V. Deoliveira, H.C. Sindra, B.S. Archanjo, M.E. Mendoza, L.S.A. Carneiro, C.D. Buarque, P.M. Esteves, *ChemCatChem*. **2016**, *8*, 743–750.

[5] R.A. Maia, F. Berg, V. Ritleng, B. Louis, P.M. Esteves, *Chem. - A Eur. J.* **2020**, *26*, 2051–2059.

[6] F.L. Oliveira, A. S. França, A.M. Castro, R.O.M. Alves de Souza, P.M. Esteves, R.S.B. Gonçalves, *Chempluschem*. **2020**, *85*, 2051–2066.

[7] A.K. Mandal, J. Mahmood, J.B. Baek, *ChemNanoMat*. **2017**, *3*, 373–391.

[8] R. Xue, H. Guo, T. Wang, L. Gong, Y. Wang, J. Ai, D. Huang, H. Chen, W. Yang, *Anal. Methods*. **2017**, *9*, 3737–3750.

[9] S.K.S. Freitas, R.S. Borges, C. Merlini, G.M.O. Barra, P.M. Esteves, *J. Phys. Chem. C*. **2017**, *121*, 27247–27252.

[10] Y. Zeng, R. Zou, Y. Zhao, *Adv. Mater.* **2016**, *28*, 2855–2873.

[11] J. Ozdemir, I. Mosleh, M. Abolhassani, L.F. Greenlee, R.R. Beitle, M.H. Beyzavi, *Front. Energy Res.* **2019**, *7*, 77.

[12] P.J. Waller, F. Gándara, O.M. Yaghi, *Acc. Chem. Res.* **2015**, *48*, 3053–3063.

[13] D. Ongari, A. V. Yakutovich, L. Talirz, B. Smit, *ACS Cent. Sci.* **2019**, *5*, 1663–1675.

[14] T.F. Willems, C.H. Rycroft, M. Kazi, J.C. Meza, M. Haranczyk, *Microporous Mesoporous Mater.* **2012**, *149*, 134–141.