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

Felipe Lopes de Oliveira, Federal University of Rio de Janeiro

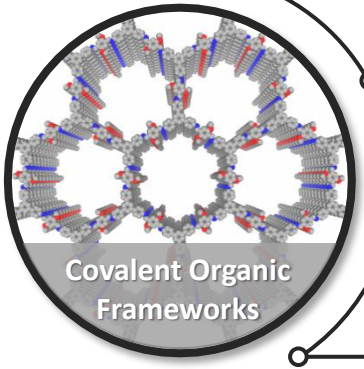
UFRJ

Michael Nazarkovsky, Pontifical Catholic University of Rio de Janeiro

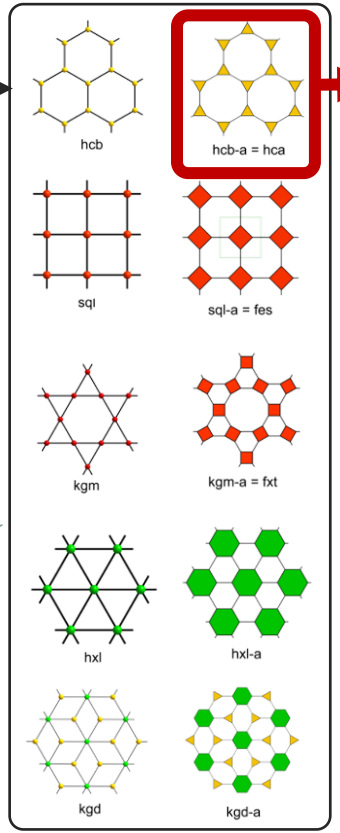
PUC RIO

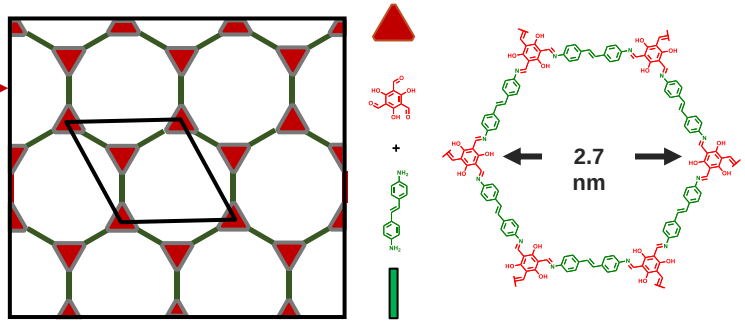
PART 1.
Introduction

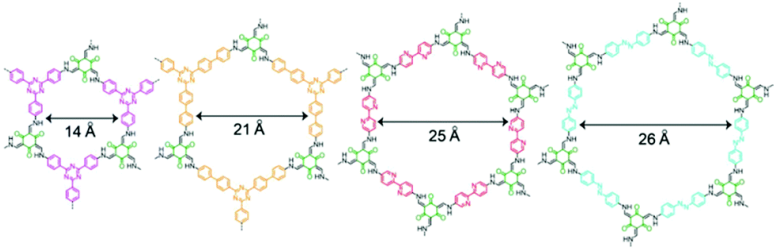
Covalent Organic Frameworks (COFs)

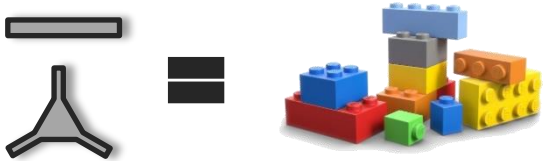


- Organic
- Crystalline
- Nanoporous
- Stable
- Reticular

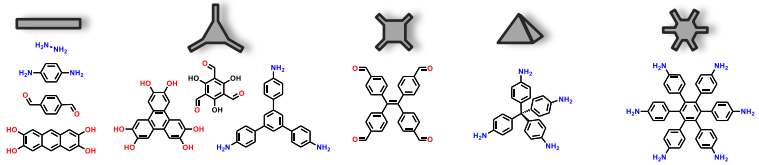




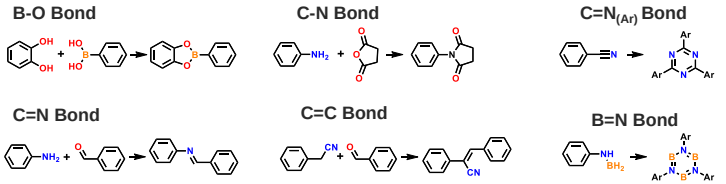




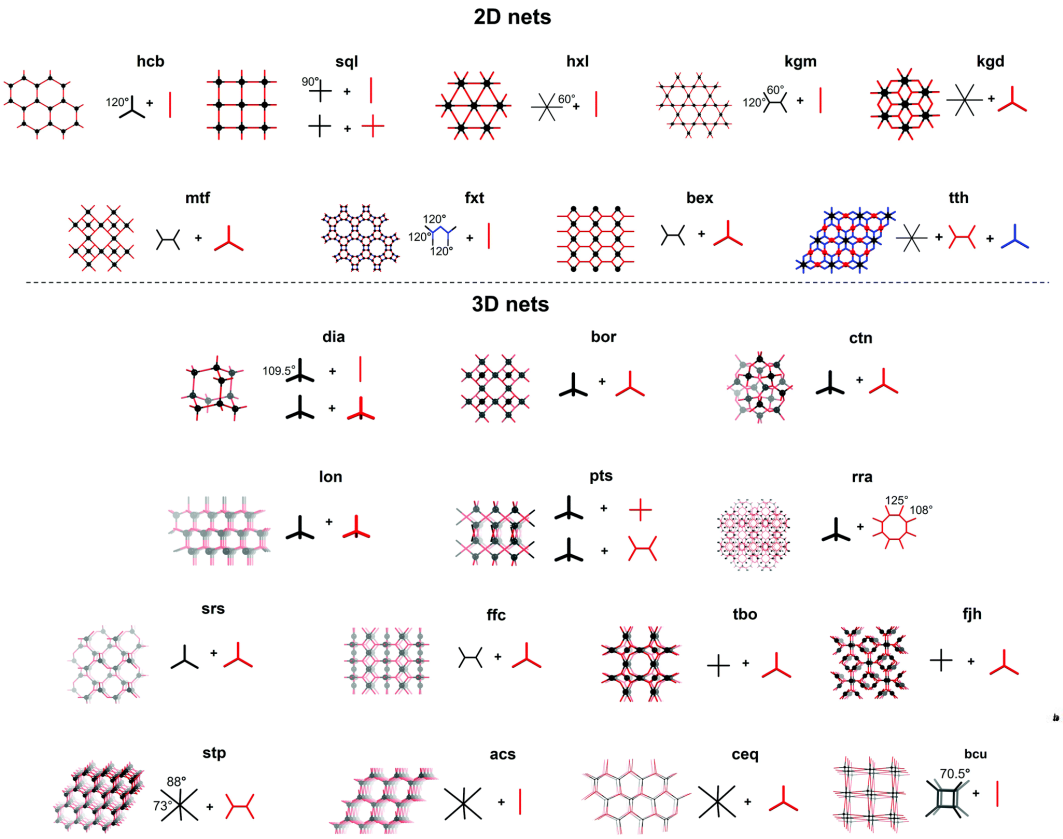
Different Building blocks tooplogy

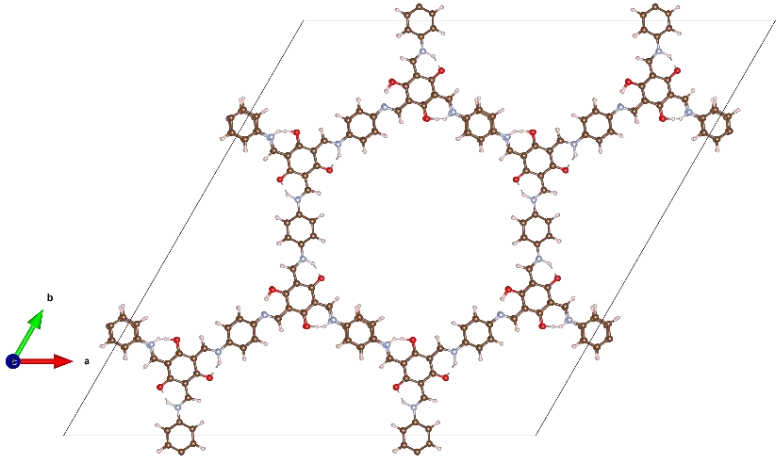


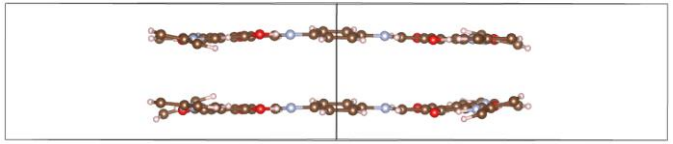
Different bonds types



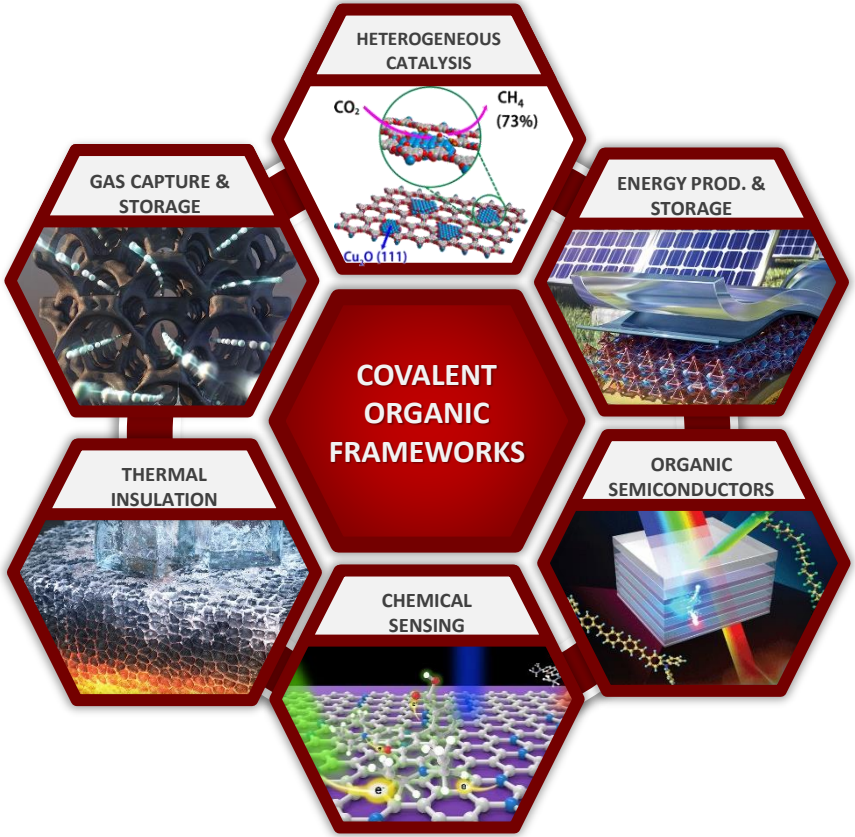
Covalent Organic Frameworks (COFs)







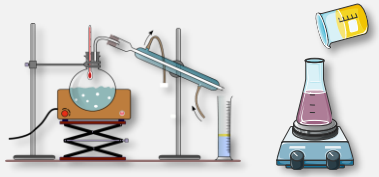
Application of Covalent Organic Frameworks



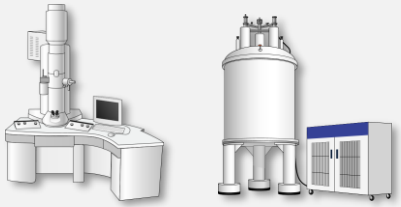
Classical paradigm of materials discovery

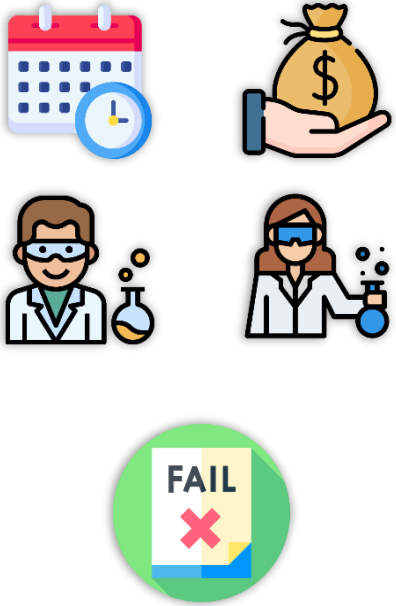
CHEMICAL INTUITION + AVAILABLY MOLECULES

SYNTHESIS

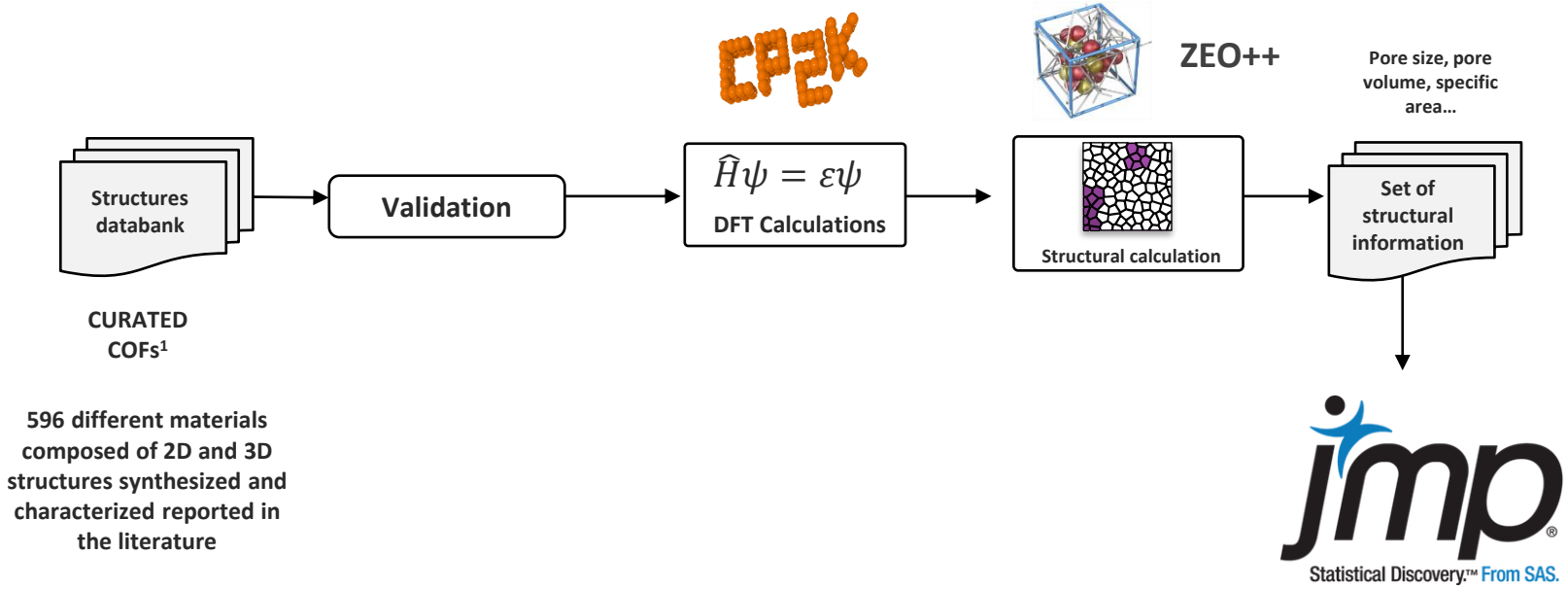


CHARACTERIZATION





Our computational approach



1. ONGARI, Daniele et al. Building a consistent and reproducible database for adsorption evaluation in covalent–organic frameworks. **ACS central science**, v. 5, n. 10, p. 1663-1675, 2019.

Some important properties

COF_ID – 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 m^2/cm^3

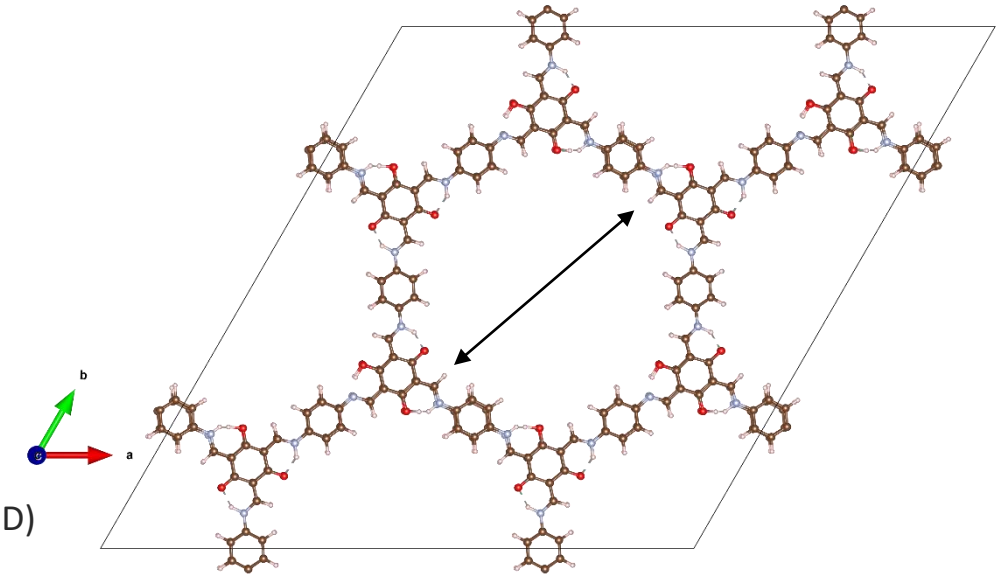
AVF - Accessible volume fraction

Vp - Pore volume cm^3/g

N(chan) - Number of accessible channels

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

ChanSize - Channel size [Å]



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

Felipe Lopes de Oliveira, Federal University of Rio de Janeiro

UFRJ

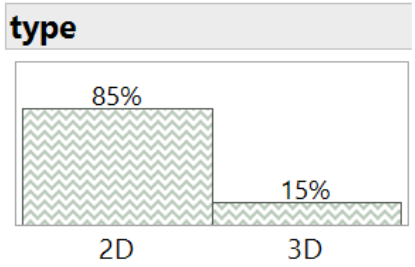
Michael Nazarkovsky, Pontifical Catholic University of Rio de Janeiro

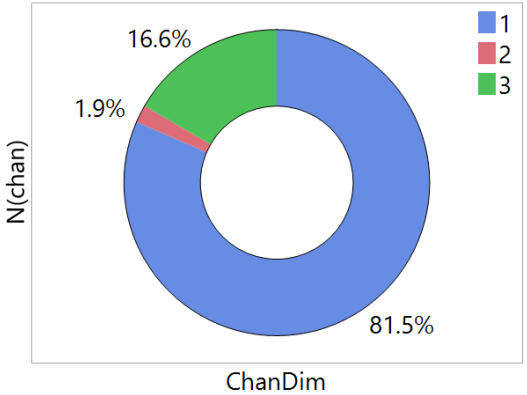
PUC RIO

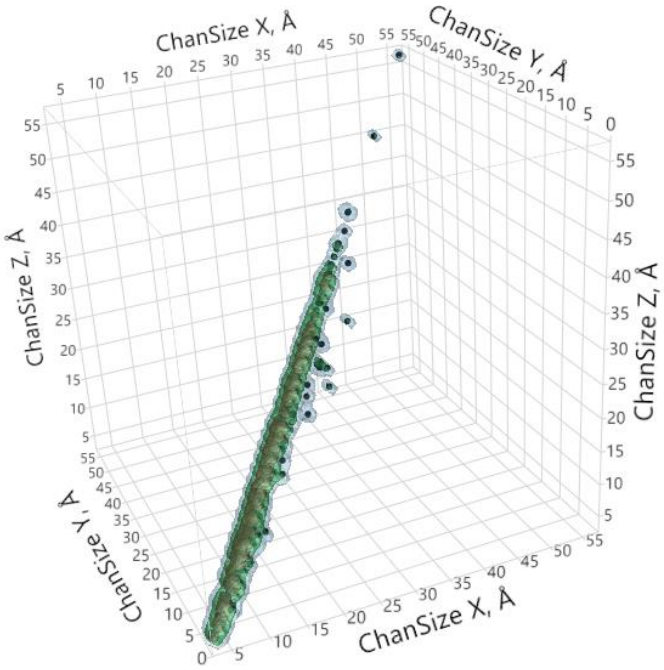
PART 2.
Data Pretreatment,
Multivariate Analysis,
Unsupervised and Supervised Predictive Modeling

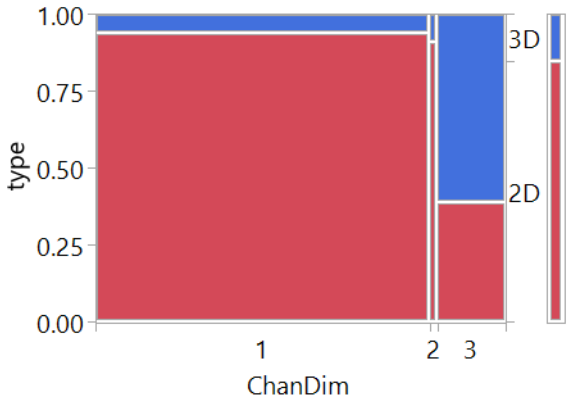
Data Preparation and Preliminary Visualization

- **Exlusion of 6 erroneous structures containg “0” values for channels and textural properties;**
- **Ordering in 10 continuous and 2 nominal inputs with a binary nominal output “type” (85% of 2D and 15% of 3D structures);**
- **No missing values.**



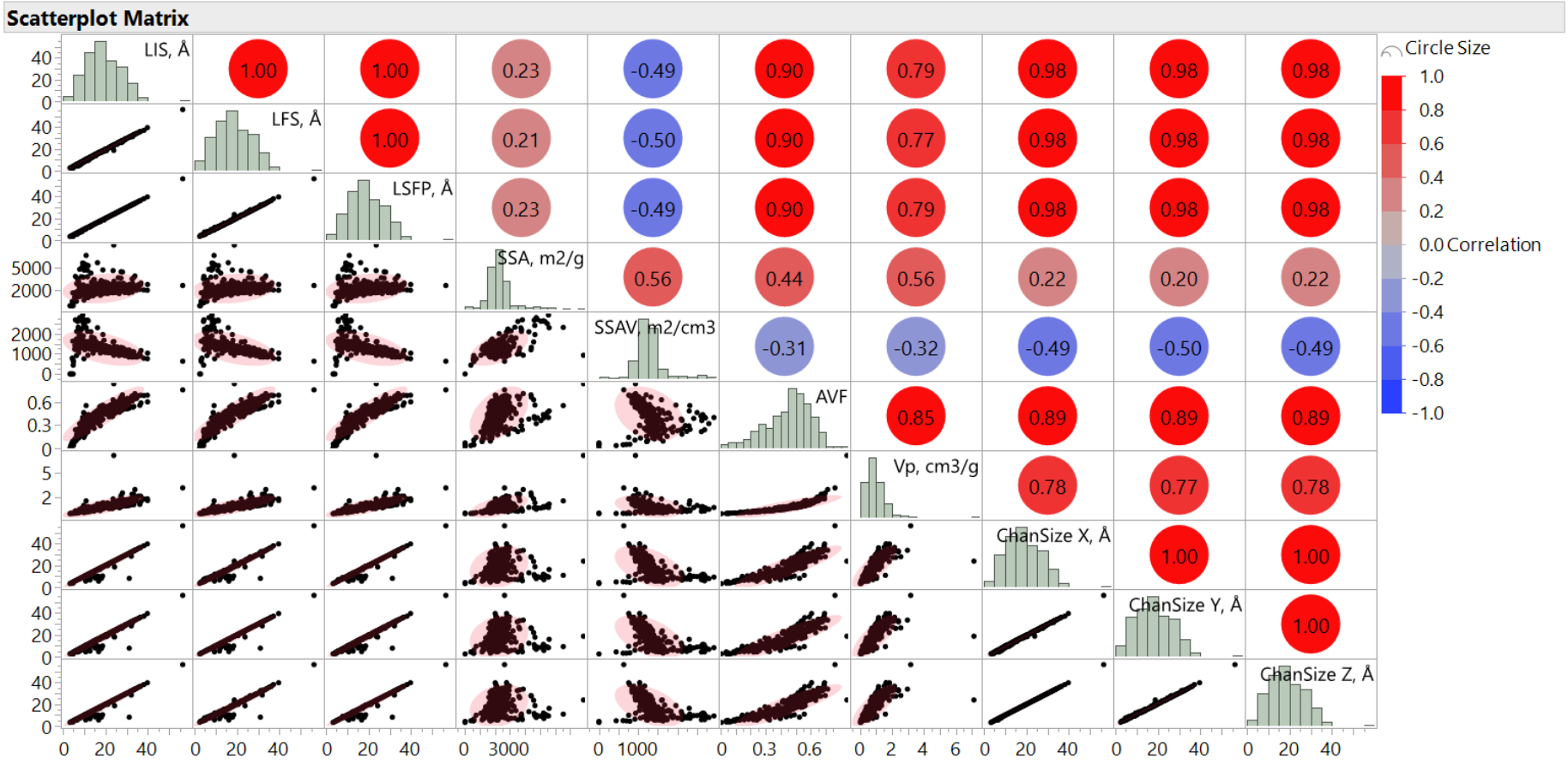




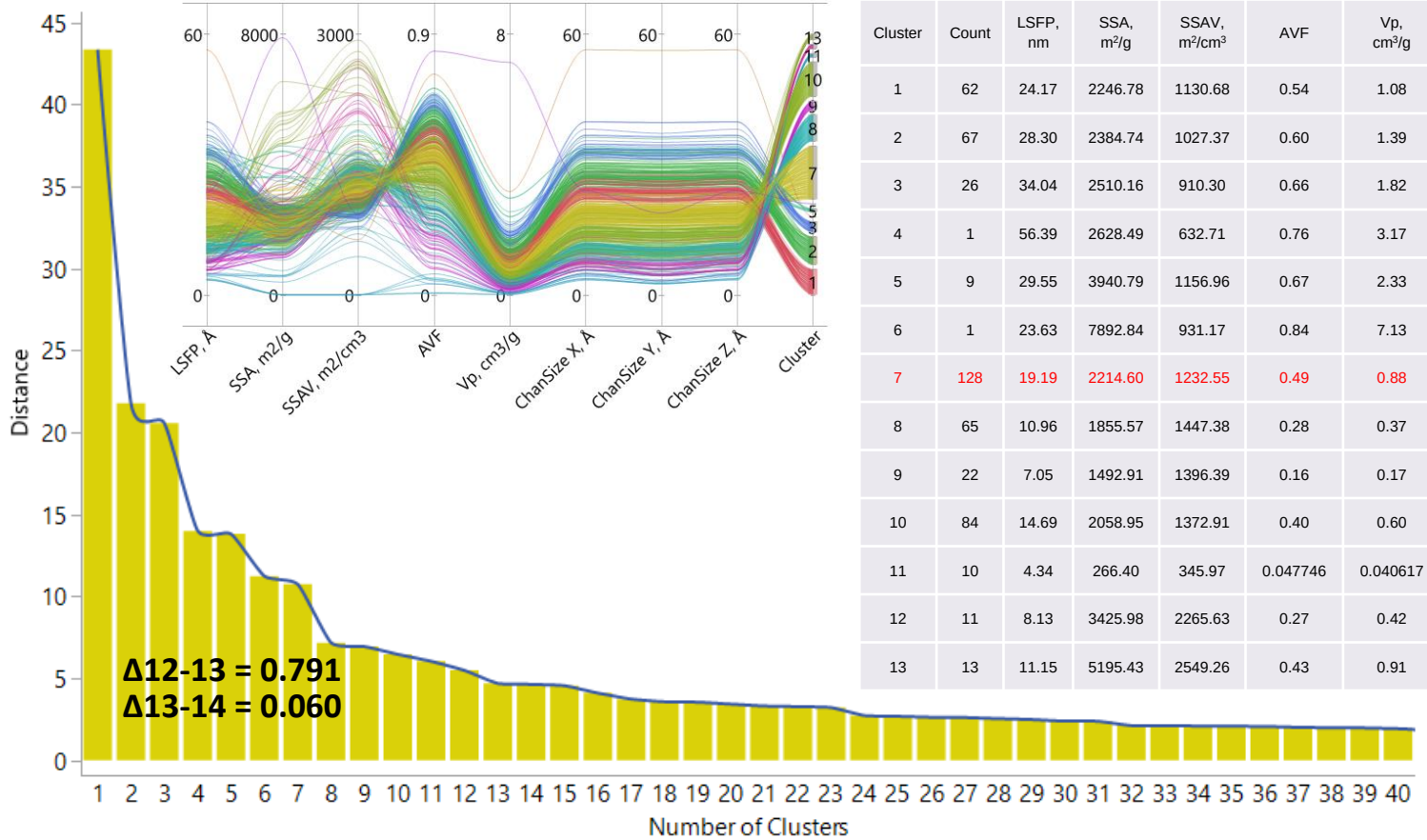


Portion %	2D	3D	Total
1	93.96	6.04	81.49
2	90.91	9.09	1.87
3	38.78	61.22	16.64
Total, %	84.72	15.28	100

Multivariate Analysis – 2D



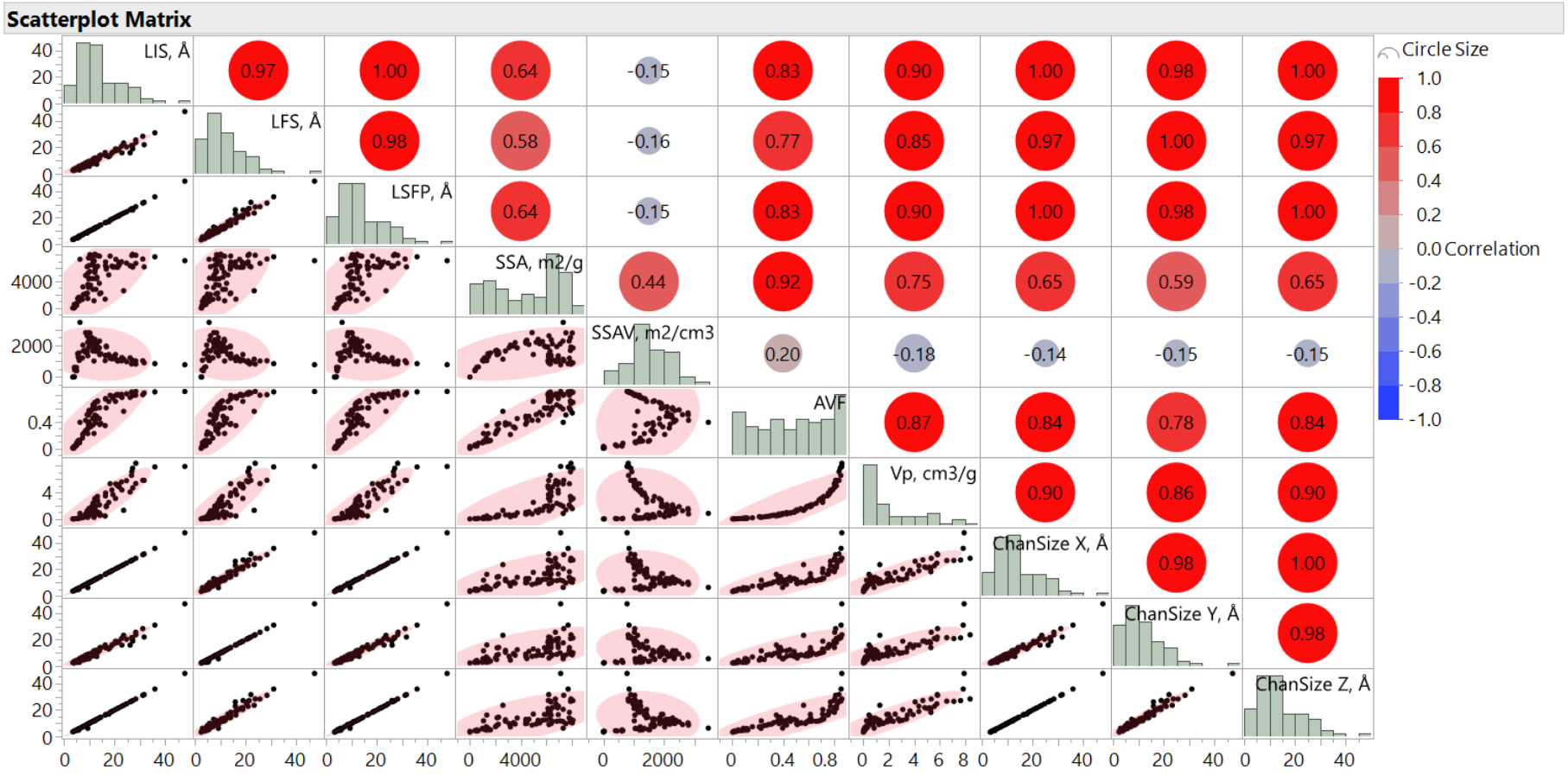
Hierarchical Clustering for 2D structures



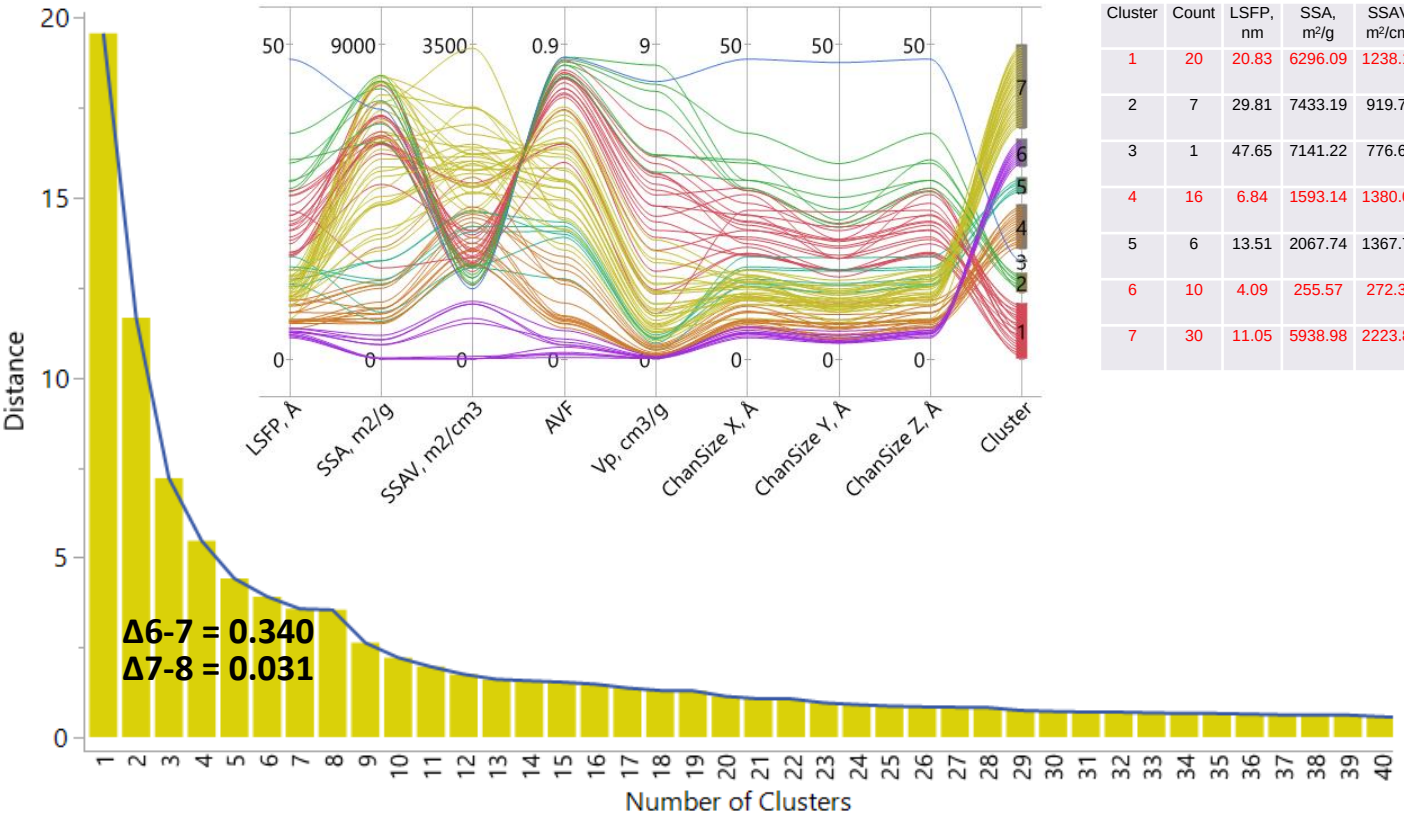
Cluster	Count	LSFP, nm	SSA, m^2/g	SSAV, m^2/cm^3	AVF	Vp, cm^3/g	ChanSize X	ChanSize Y	ChanSize Z
1	62	24.17	2246.78	1130.68	0.54	1.08	24.15	23.71	24.15
2	67	28.30	2384.74	1027.37	0.60	1.39	28.24	27.87	28.24
3	26	34.04	2510.16	910.30	0.66	1.82	34.05	33.67	34.05
4	1	56.39	2628.49	632.71	0.76	3.17	56.40	56.25	56.40
5	9	29.55	3940.79	1156.96	0.67	2.33	29.56	28.78	29.56
6	1	23.63	7892.84	931.17	0.84	7.13	23.92	18.76	23.92
7	128	19.19	2214.60	1232.55	0.49	0.88	19.13	18.65	19.13
8	65	10.96	1855.57	1447.38	0.28	0.37	10.29	9.42	10.27
9	22	7.05	1492.91	1396.39	0.16	0.17	7.13	6.07	7.05
10	84	14.69	2058.95	1372.91	0.40	0.60	13.91	13.36	13.91
11	10	4.34	266.40	345.97	0.047746	0.040617	4.38	3.31	4.38
12	11	8.13	3425.98	2265.63	0.27	0.42	8.32	6.68	8.11
13	13	11.15	5195.43	2549.26	0.43	0.91	11.00	9.96	10.94

Shown 40 of 498

Multivariate Analysis – 3D



Hierarchical Clustering for 3D structures



Cluster	Count	LSFP, nm	SSA, m^2/g	SSAV, m^2/cm^3	AVF	Vp, cm^3/g	ChanSize X	ChanSize Y	ChanSize Z
1	20	20.83	6296.09	1238.13	0.76	4.14	21.01	17.22	20.96
2	7	29.81	7433.19	919.70	0.84	6.86	29.89	24.72	29.86
3	1	47.65	7141.22	776.63	0.86	7.93	47.65	47.10	47.65
4	16	6.84	1593.14	1380.00	0.16	0.19	6.54	5.20	6.52
5	6	13.51	2067.74	1367.75	0.34	0.52	13.50	13.01	13.50
6	10	4.09	255.57	272.35	0.034086	0.032363	4.17	3.00	4.11
7	30	11.05	5938.98	2223.87	0.53	1.52	11.16	8.92	11.11

Shown 40 of 89

Portion of total variation in each column absorbed by clustering

2D

Column Summary

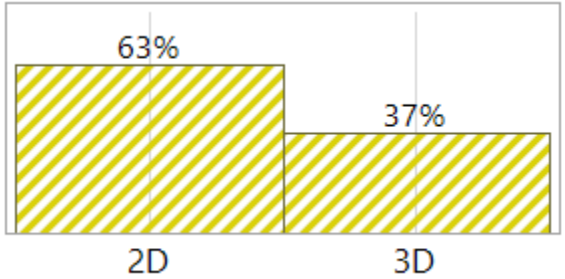
Column	RSquare	.2.4.6.8
ChanSize X, Å	0.9421	
ChanSize Z, Å	0.9420	
ChanSize Y, Å	0.9405	
LSFP, Å	0.9325	
AVF	0.9289	
Vp, cm3/g	0.9201	
SSAV, m2/cm3	0.8463	
SSA, m2/g	0.8010	

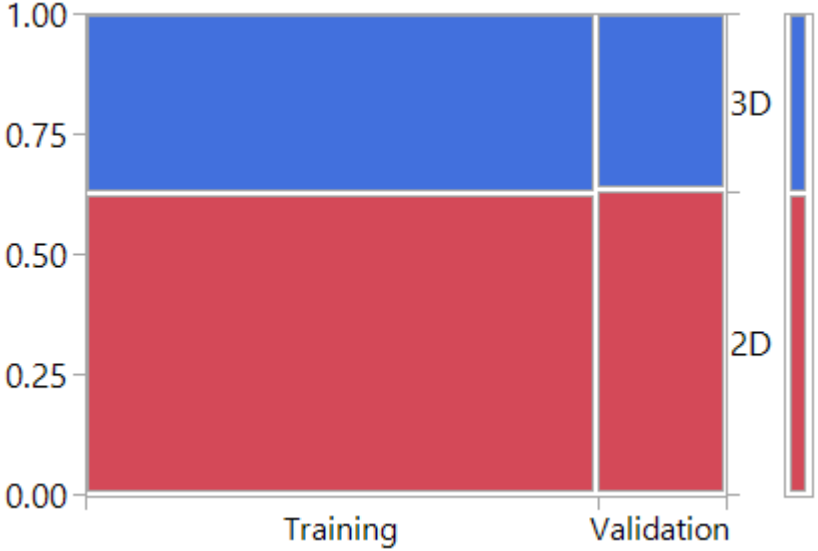
3D

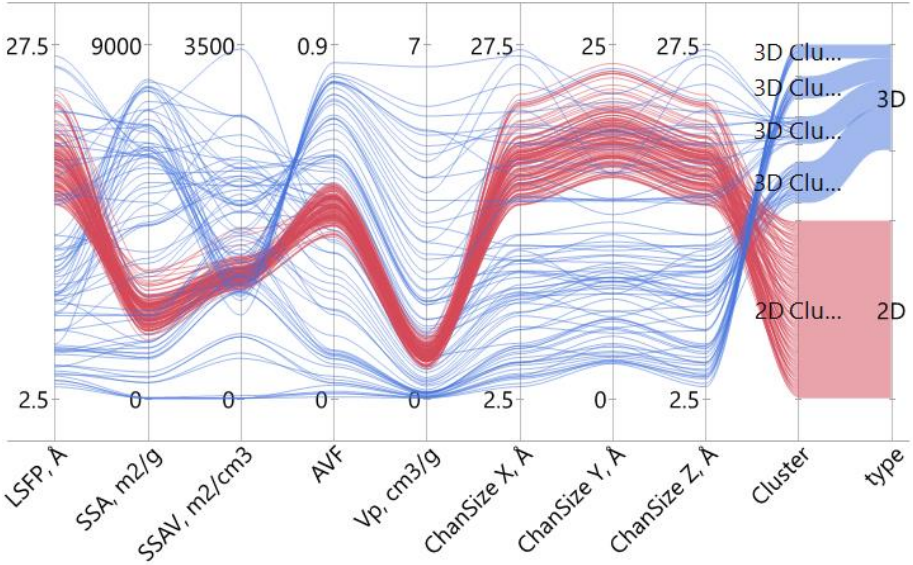
Column Summary

Column	RSquare	.2.4.6.8
ChanSize X, Å	0.9332	
LSFP, Å	0.9322	
ChanSize Z, Å	0.9319	
ChanSize Y, Å	0.9200	
AVF	0.9128	
Vp, cm3/g	0.8600	
SSA, m2/g	0.8387	
SSAV, m2/cm3	0.8108	

2D/3D by the most massive clusters







Count Total %	2D	3D	Total
Training	102 50.00	61 29.90	163 79.90
Validation	26 12.75	15 7.35	41 20.10
Total	128 62.75	76 37.25	204 100

Logistic Regression

Whole Model Test

Model	-LogLikelihood	DF	χ^2	$p>\chi^2$
Difference	94.00774	2	188.0155	<.0001*
Full	13.76301			
Reduced	107.77075			

Variable	LogWorth	p-Value
Vp, cm3/g	41.633	0.00000
AVF	37.202	0.00000

Metrics	Training	Validation
Entropy R^2	0.8723	0.9521
Generalized R^2	0.9332	0.9761
Mean –Log *p*	0.0844	0.0315
Misclassification Rate	0.0429	0.0000
N	163	41

Term	Estimate	Std Error	χ^2	$p>\chi^2$
Intercept	-36.076807	11.01415	10.73	0.0011*
AVF	140.878677	36.651292	14.77	0.0001*
Vp, cm3/g	-31.147424	7.4543375	17.46	<.0001*

Confusion Matrix

Training

Actual type	Predicted Count 2D	3D
2D	99	3
3D	4	57

Validation

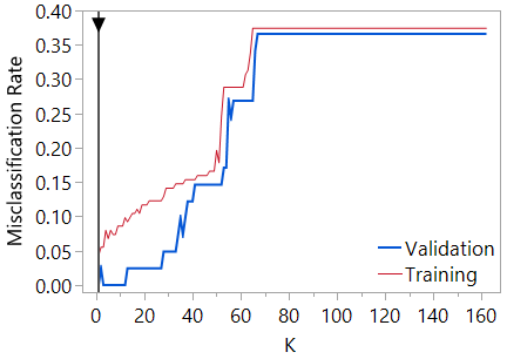
Actual type	Predicted Count 2D	3D
2D	26	0
3D	0	15

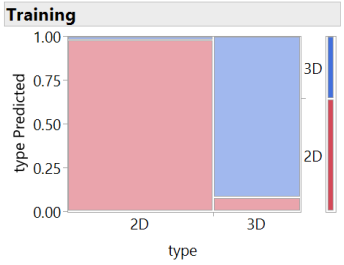
-36.0768067704061 + 140.878676823079AVF - 31.147423524102Vp = Lin2D

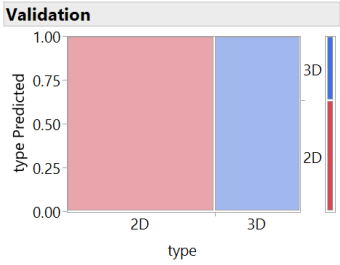
1/(1 + exp(-Lin2D)) = Probability_{2D}

1/(1 + exp(Lin2D)) = Probability_{3D}

K Nearest Neighbors (Euclidean distance. K_{max} = 162, K_{best} = 1)







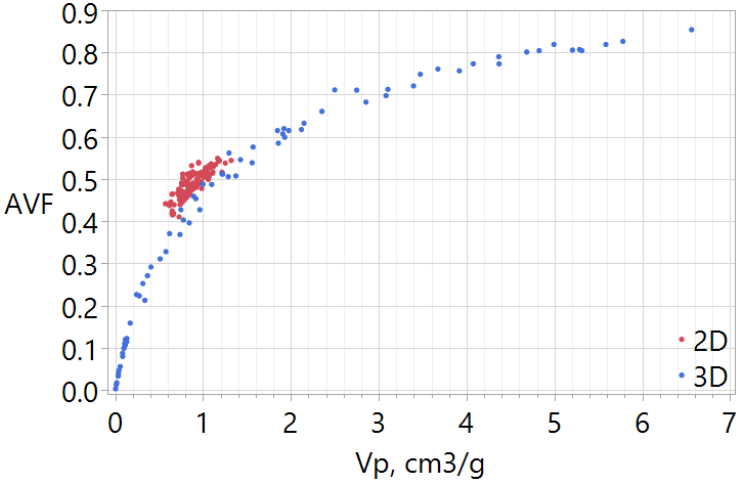
Confusion Matrix for Best K=1

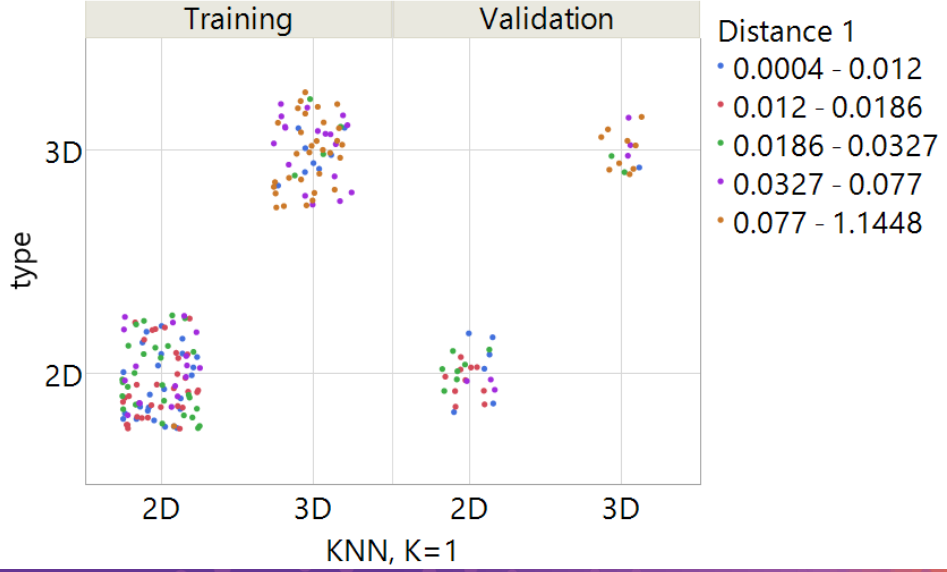
Training

Actual type	Predicted Count 2D	3D
2D	100	2
3D	5	56

Validation

Actual type	Predicted Count 2D	3D
2D	26	0
3D	0	15





Quadratic Discriminant Analysis

Score Summaries

Source	Count	Number Misclassified	Percent Misclassified	Entropy RSquare	-2LogLikelihood
Training	163	8	4.90798	0.80767	41.4546
Validation	41	0	0.00000	0.98341	

Group Means

Count	type	AVF	Vp, cm3/g
102	2D	0.48692431	0.8707681
61	3D	0.44836164	1.7041224

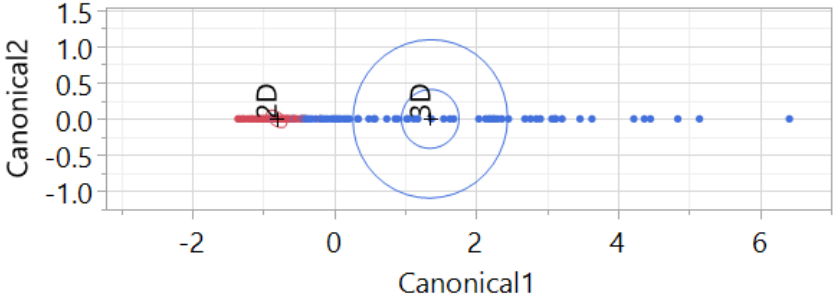
Training

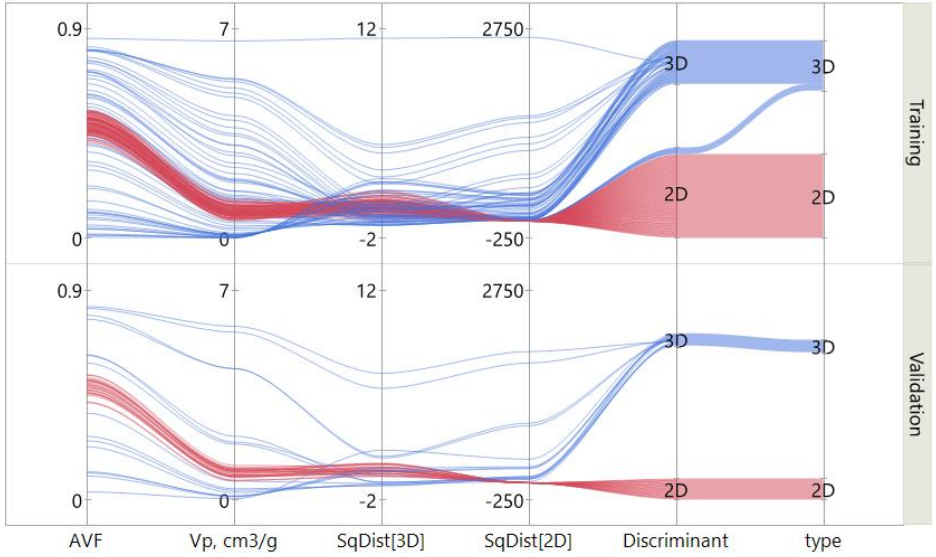
Actual type	Predicted Count 2D	Predicted Count 3D
2D	102	0
3D	8	53

Validation

Actual type	Predicted Count 2D	Predicted Count 3D
2D	26	0
3D	0	15

Canonical Plot





Naive Bayes

Metrics	Training	Validation
Entropy R^2	0.6355	0.9744
Generalized R^2	0.7750	0.9874
Mean -Log *p*	0.2410	0.0168
Misclassification Rate	0.0736	0.0000
N	163	41

Confusion Matrix

Training

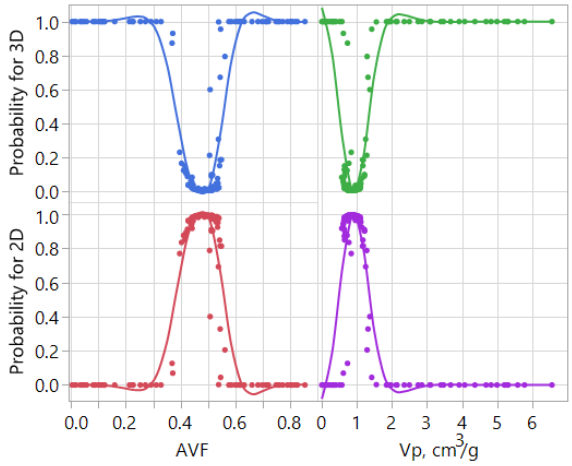
Actual type	Predicted Count 2D	3D
2D	101	1
3D	11	50

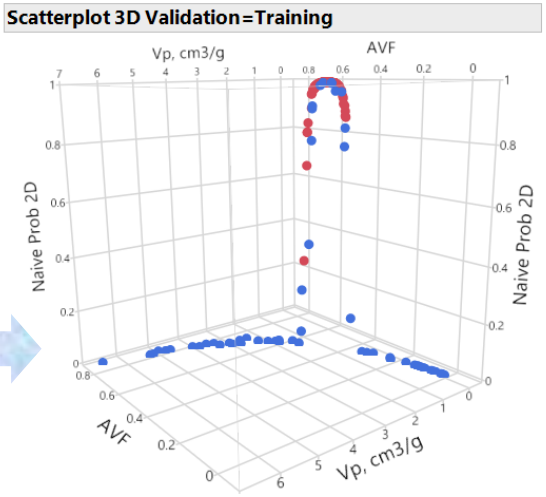
Validation

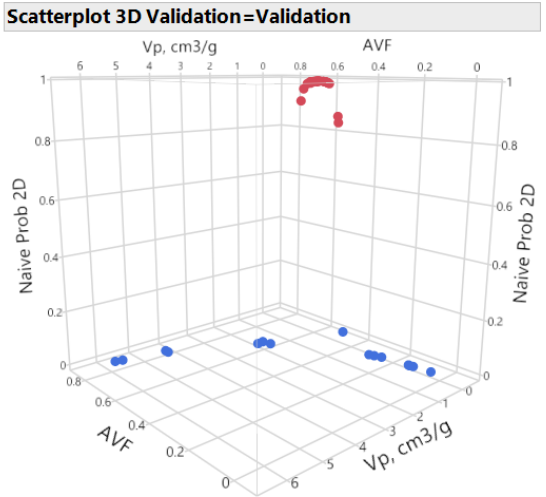
Actual type	Predicted Count 2D	3D
2D	26	0
3D	0	15

Independent Inputs (Overall)

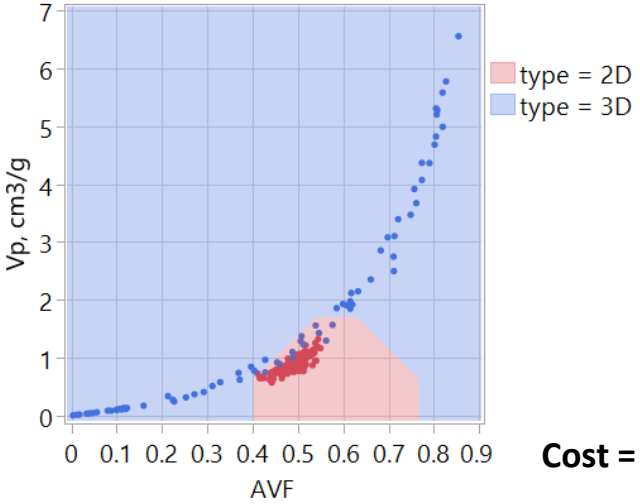
Column	Main Effect	Total Effect
Vp, cm^3/g	0.217	0.879
AVF	0.122	0.783

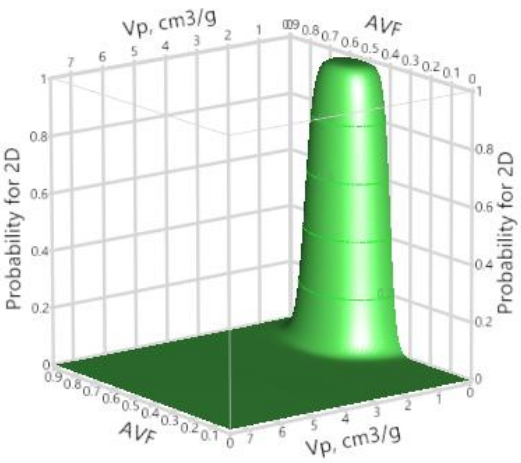


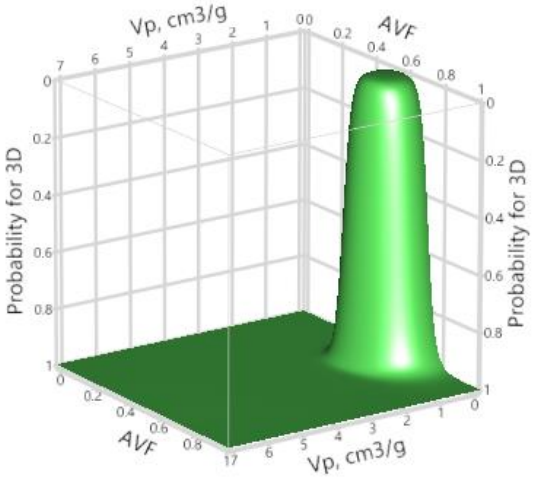




Support Vector Machine

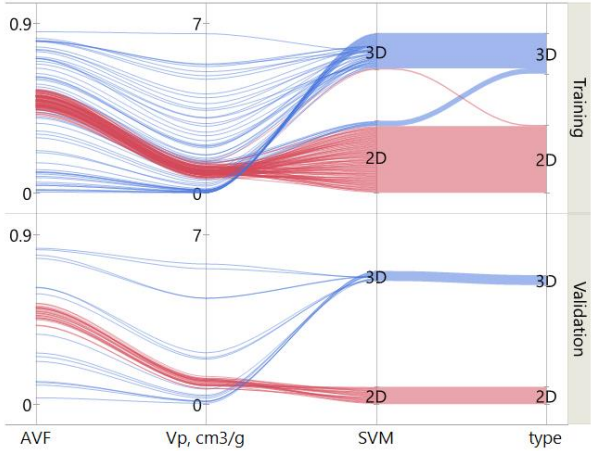






Cost = 1, Gamma = 0.5, #SV = 46

Metrics	Training	Validation
Entropy R^2	0.7849	0.9097
Generalized R^2	0.8805	0.9537
Mean -Log *p*	0.1422	0.0593
Misclassification Rate	0.0613	0.0000
N	163	41



Confusion Matrix

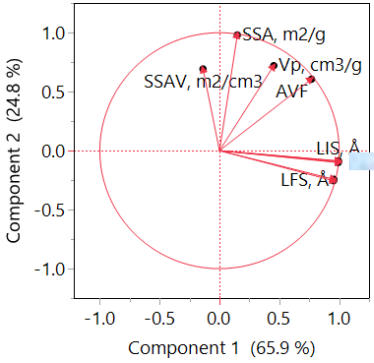
Training

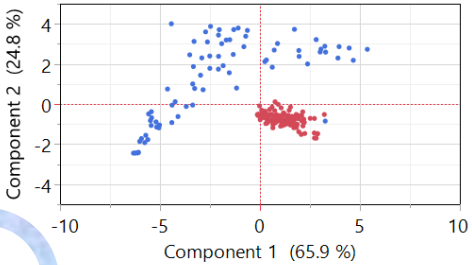
Actual type	Predicted Count 2D	Predicted Count 3D
2D	100	2
3D	8	53

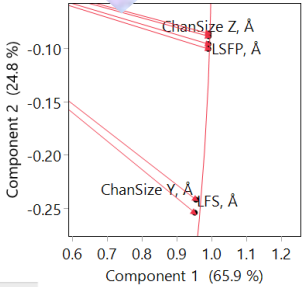
Validation

Actual type	Predicted Count 2D	Predicted Count 3D
2D	26	0
3D	0	15

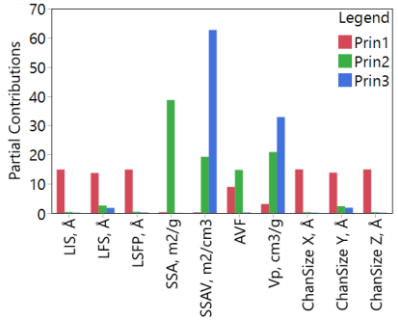
PCA……







Plot of Partial Contribution of Variables



Eigenvalues

Number	Eigenvalue	Percent	20406080	Cum Percent
1	6.5938	65.938		65.938
2	2.4836	24.836		90.774
3	0.7914	7.914		98.688
4	0.0584	0.584		99.272
5	0.0351	0.351		99.623
6	0.0250	0.250		99.873
7	0.0120	0.120		99.992
8	0.0005	0.005		99.997
9	0.0002	0.002		99.999
10	0.0001	0.001		100.000

…and PCA based Logistic Regression

Whole Model Test

Model	-LogLikelihood	DF	χ^2	$p>\chi^2$
Difference	93.99795	2	187.9959	<.0001*
Full	13.77280			
Reduced	107.77075			

Metrics	Training	Validation
Entropy R^2	0.8722	0.9201
Generalized R^2	0.9331	0.9593
Mean –Log *p*	0.0845	0.0525
Misclassification Rate	0.0061	0.0000
N	163	41

Confusion Matrix

Training

Actual type	Predicted Count 2D	Predicted Count 3D
2D	102	0
3D	1	60

Validation

Actual type	Predicted Count 2D	Predicted Count 3D
2D	26	0
3D	0	15

Term	Estimate	Std Error	χ^2	$p>\chi^2$
Intercept	-0.1203949	0.6012218	0.04	0.8413
Prin1	1.80736932	0.3911632	21.35	<.0001*
Prin2	-4.004678	0.934766	18.35	<.0001*

-0.120394878783972 + 1.80736932265296Prin1 - 4.00467801810539Prin2 = Lin2D

1/(1 + exp(-Lin2D)) = Probability_{2D}

1/(1 + exp(Lin2D)) = Probability_{3D}

Conclusions

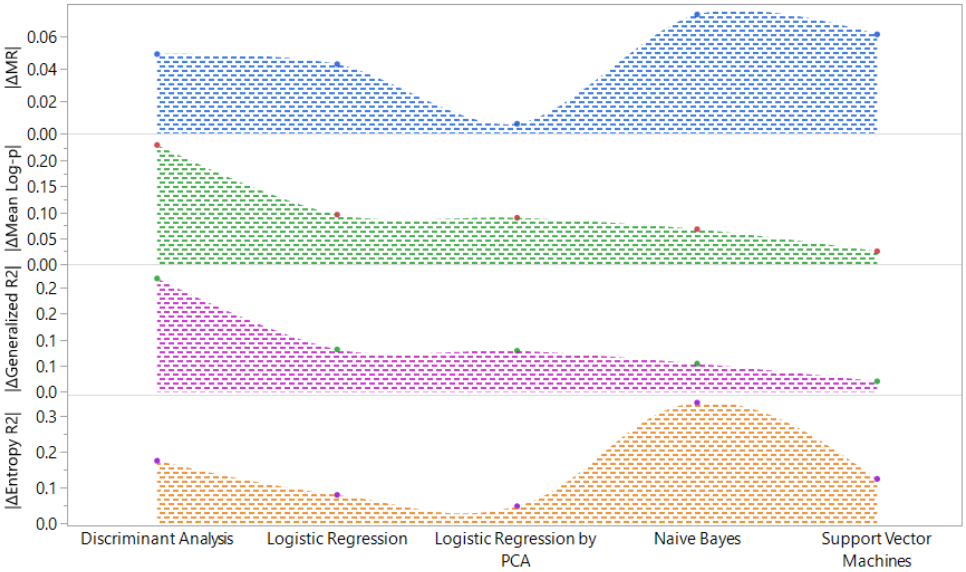
Training

Model	Entropy R^2	Generalized R^2	Mean -Log p	MR	N
Logistic Regression	0.8723	0.9332	0.0844	0.0429	163
! Logistic Regression by PCA	0.8722	0.9331	0.0845	0.0061	163
Discriminant Analysis	0.8077	0.8948	0.1272	0.0491	163
Support Vector Machines	0.7849	0.8805	0.1422	0.0613	163
Naive Bayes	0.6355	0.7750	0.241	0.0736	163

Validation

Model	Entropy R^2	Generalized R^2	Mean -Log p	MR	N
Discriminant Analysis	0.9834	0.9919	0.0109	0.0000	41
Naive Bayes	0.9744	0.9874	0.0168	0.0000	41
Logistic Regression	0.9521	0.9761	0.0315	0.0000	41
Logistic Regression by PCA	0.9201	0.9593	0.0525	0.0000	41
Support Vector Machines	0.9097	0.9537	0.0593	0.0000	41

What's the best of the best?



Thank You!

Obrigado!

Дякую!

Acknowledgements:
Dr. David Kirmayer
The Hebrew University of Jerusalem - HUJI, School of Pharmacy
Jerusalem, Israel