Analysis of Aggregate Growth Behaviour on Calixarene Compounds Developed under Varying Synthesis Conditions

Alternate title: *Prediction of Surface Aggregate Properties of Calixarene Compounds using Theoretical Descriptors and Synthesis Conditions*

This report summarizes the analysis and investigation performed on the behaviour of aggregates formed over calixarene films. Calixarene compounds were synthesized under various combinations of source compounds, solvent, concentration and respective temperatures resulting in numerous growth patterns of aggregates (i.e. bumps). These aggregates are crucial to future use of the compounds themselves, as they improve the binding/carrier properties resulting in improve efficiency when transporting materials (i.e. w.r.t drug delivery systems). Using Atomic Force Microscopy (AFM), the synthesized films were scanned and processed to extract descriptive properties of the aggregates i.e. mean particle size (MPS), Polydispersity (PD), ratio of area covered by sphere to substance (RAS). From the data extracted, theoretical descriptors generated from RD-KIT toolkit were used in regression models to estimate and classify films as per their aggregate behaviour. Our results extends prior analysis performed via correlation analysis and mutual information metrics. The current observation aims to provide ground work to analyze larger sample sets of data and further create models to predict surface properties, given the prior synthesis conditions.

The compounds used in this work are:

- **$C_{46}H_{58}O_6S_4$** – Amphiphilic tert-butylthiacalix[4]mono-crown-4 – (Compound 1)
- **$C_{40}H_{40}N_4O_{18}S_4$** – Bolaamphiphilic 1,3-alternate nitrothiacalix[4]bis-crown-5 (Compound 2)

while Solvents:

- **$CHCl_3$** – Chloroform
- **$C_6H_5CH_3$** – Toluene

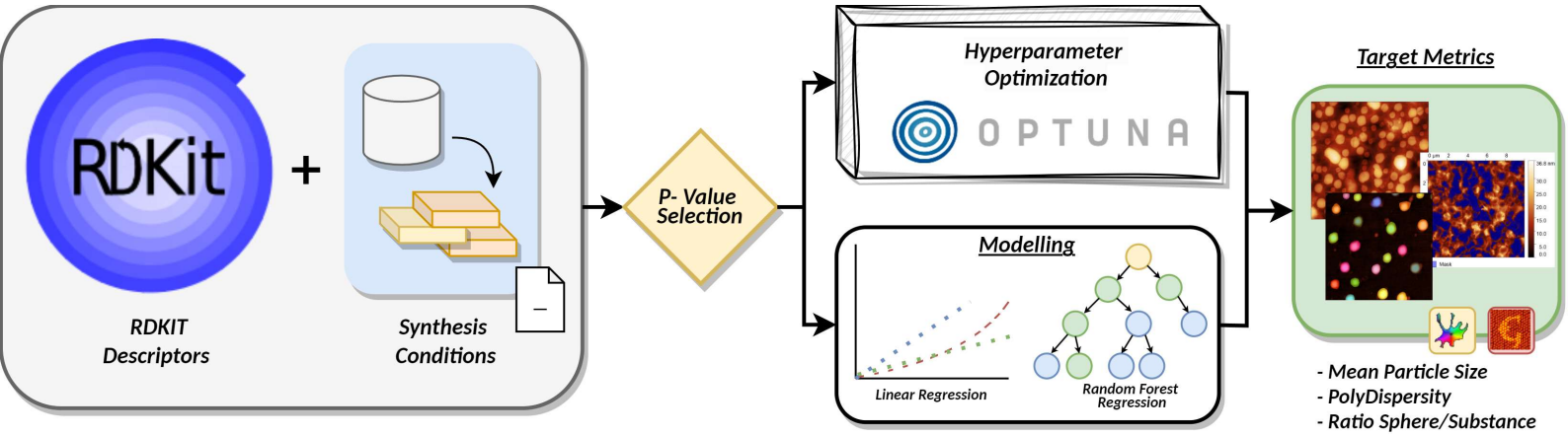


Figure 1: Pipeline of aggregate analysis and regression modelling performed using compound synthesis data and RD Kit descriptors.

Methodology

Data Collection & Processing

Table 1: An overview of synthesised films across all conditions analysed during our experimental analysis

Compound Type	*Compound Temp*	*Solvent type*	*Solvent Temp*	*Concentration of Solution*
156	23°C	***$CHCL_3$***	23°C	10^{-4}
	23°C		4°C	10^{-4}
	4°C		23°C	10^{-4}
	23°C		23°C	10^{-5}
	23°C	***Toluene***	23°C	10^{-4}
	23°C		23°C	10^{-5}
159	23°C	***$CHCL_3$***	23°C	10^{-4}
	4°C		23°C	10^{-4}
	23°C		4°C	10^{-4}
	23°C		23°C	10^{-5}
	23°C	***Toluene***	23°C	10^{-4}
	23°C		23°C	10^{-5}

Synthesis Data:

The synthesis data of calixarene compounds consists of five parameters:

1. Compounds type ($C_{46}H_{58}O_6S_4$ or $C_{40}H_{40}N_4O_{18}S_4$)
2. Solvent Type ($CHCl_3$ **or** $C_6H_5CH_3$)
3. Compound Temperature (**23°C** or **4°C**)
4. Solvent Temperature (**23°C** or **4°C**)
5. Concentration of Compound in Solution (10^{-4} or 10^{-5})

Within our regression flow, we consider this data as source information for modelling. In future versions our goal is to use above parameters as input to a model that predicts the possibility of occurrence and features of aggregates on the surface.

RD – Kit Descriptors:

In prior work, we used threshold-ed descriptors extracted from SMILES data of the compounds. From community [Cite] and our own analysis, we observed that there exists loss of information when using descriptors generated via SMILES. Hence, we have extracted current descriptors using (.SDF) files that embed spatial information of the molecular structure. Using a custom Python Script, we generate descriptors using the RD-KIT toolbox using Spatial Dimensional Files for both compounds and Solvents. This results in a total of 117 descriptors after removal of null or zero fields.

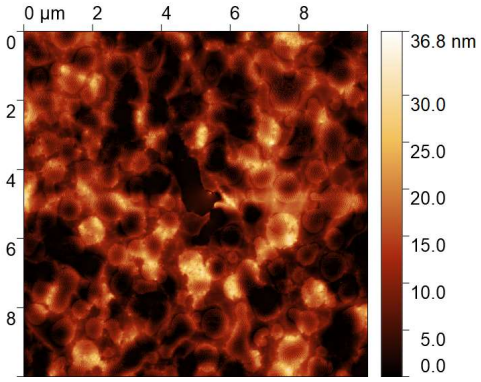
Extraction of Aggregate Features:

From the acquired AFM scanned images. We extracted descriptive aggregate properties:

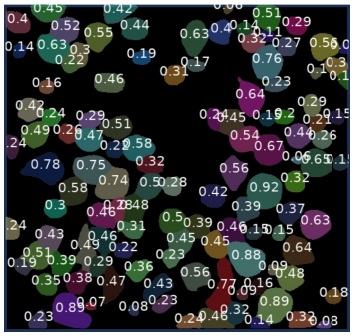
1. **Mean Particle Size (MPS)**: The diameter of spheres from each sample were measured individually using an external toolkit (Input-from-Anna) and the mean of size of spheres considered as the respective MPS for the film

2. **Polydispersity (PD)**: (**Change polydispersity to just Dispersity?**) We define polydispersity as the heterogeneity of the size of aggregates on the films. It is calculated as the ratio of standard devition of the diameter of each aggregate in a film to the calculated MPS. (i.e. a larger PD represents a sample with greater variance of the size of aggregates whereas a low PD score represents a sample with uniform size of aggregates)

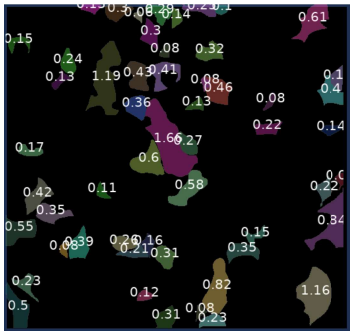
3. **Ratio of Sphere to Substance**: (Change to **Ratio of Aggregate Area**?)
   The metric defines the area covered by the aggregates relative the amount of area covered by substrate (i.e. the substance itself). We use a combination of Cellpose Plus (Citation) and Gwyddion's Shade Data tool (citation), to distinguish the aggregates from the substance and classify extract their area.



Source sample



Area of Spheres = 44.24цm^2



Area of Substrate = 18.38цm^2

Sample #	*Mean Particle Size*	*Poly Dispersity*	*Ratio of Aggregates*
1. Comp 1	566.36	0.3	0.44
2. Comp 1	538.82	0.37	0.44
3. Comp 1	221.96	0.32	0.12
4. Comp 1	272.23	0.56	0.33
5. Comp 1	551.93	0.39	0.45
6. Comp 1	262.36	0.39	0.04

7. Comp 2	0	0	0
8. Comp 2	423.37	0.35	0.43
9. Comp 2	333.66	0.66	0.25
10. Comp 2	290.41	0.27	0.26
11. Comp 2	703.28	0.14	0.25
12. Comp 2	0	0	0

Table 2: The calculated descriptive metrics for Compounds 1 and 2

Results

Regression:

We perform regression using in three perspective cases to ensure complete coverage and covering for chance occurences i.e. regression using:
1. Exclusively **Theoretical Descriptors (**from **RD-Kit)**
2. Exclusively **Synthesis conditions**
3. Both **Theoretical descriptors** & **Synthesis Conditions**

aimed at prediting:

- **Mean Particle Size (MPS)**
- **Polydispersity (PD)**
- **Ratio of Aggregate area**

Linear Regression

Case (1): Prediction using **RD KIT metrics**

Predictor	*Target*	r^2	*Mean Absolute Error*
TPSA_C	Mean Particle Size **(MPS)**	0.009	308.00
PEOE_VSA6_C		0.009	308.00
TPSA_C PEOE_VSA6_C		0.009	308.00
Estate_VSA8_S	**Polydispersity**	-2.4	0.260
SlogP_VSA4_S		-2.4	0.260
VSA_EState2_C SlogP_VSA4_S, fr_ether, TPSA		-2.59	0.266
Estate_VSA8_C, SlogP_VSA4_S, fr_ether, TPSA		-2.59	0.266
TPSA, SlogP_VSA4_C	**Ratio Sphere Substance**	-0.63	0.217
SlogP_VSA4_C, NumAliphaticHeterocycles, TPSA		-0.63	0.217
SlogP_VSA4_C, NumAliphaticRings, TPSA		-0.63	0.217
SlogP_VSA4_C		-0.63	0.217
SlogP_VSA4_C EState_VSA2		-0.63	0.217
Estate_VSA10 NumAliphaticHeterocycles		-0.63	0.217

Case (2): Prediction using **Synthesis Conditions**

Predictor	***Target***	***r^2***	***Mean Absolute Error***
Compound Type Temp Solvent Hydophilic_sphericity Ratio Area	Mean Particle Size **(MPS)**	0.654	143.87
Temp – Compound Temp – Solvents Ratio Area		0.640	179.318
Temp – Compound Temp – Solvents Hydophilic_sphericity Ratio Area		0.633	163.37
Hydophilic_sphericity Hydrophobic_sphericity Hydrophilic_packfactor	**Polydispersity**	-1.79	0.227
Hydrophobic_sphericity full_mol_packfactor Hydrophilic_packfactor		-1.79	0.227
full_mol_packfactor Hydrophilic_packfactor		-1.80	0.229
Hydrophilic_spherecity	**Ratio Sphere Substance**	-0.620	0.216
Hydrophilic_packfactor		-0.712	0.224
Hydrophobic_packfactor		-0.888	0.237

Case (3): Prediction using **Synthesis Conditions + RD KIT**

Predictor	***Target***	***r^2***	***Mean Absolute Error***
Temp_Comp Temp_Solvent Hydrophilic_sphericity	Mean Particle Size **(MPS)**	0.433	233.929
Temp_Comp Temp_Solvent Hydrophilic_packingfact		0.425	235.35
Temp_Comp Temp_Solvent		0.401	238.5
TPSA, Sphericity Pack factor etc	**Polydispersity**	-1.79	0.227
SlogP_VSA4_S or Estate_VSA8_S (Solvent)		-2.40	0.260
Estate_VSA8_C or VSA_EState2_C (Compound)		-3.291	0.283
Hydophilic_sphericity	**Ratio Sphere Substance**	-0.620	0.216
SlogP_VSA4_C TPSA Num_Aliphatic_Rings other RDKIT merics		-0.632	0.217

Random Forest Regression

Case (1): Prediction using **RD KIT metrics**

Predictor	***Target***	***r^2***	***Mean Absolute Error***
TPSA_C	Mean Particle Size **(MPS)**	-0.017	313.33
PEOE_VSA6_C		-0.017	313.33
TPSA_C PEOE_VSA6_C		-0.121	332.67
Estate_VSA8_S SlogP_VSA4_C SlogP_VSA4_S	**Polydispersity**	-1.790	0.226
Estate_VSA8_S SlogP_VSA4_C Num_aliphatic_heterocycles Num_Aliphatic_rings		-1.790	0.226
VSA_Estate2_C Estate_VSA8_S SlogP_VSA4_C or NumAliphaticHeterocycles/ TPSA		-1.80	0.229
All_Metrics	**Ratio Sphere Substance**	-0.771	
Estate_VSA8_S SlogP_VSA4_S		-0.945	
Estate_VSA8_S SlogP_VSA4_C VSA_EState2		-2.01	

Case (2): Prediction using **Synthesis Conditions**

Predictor	***Target***	***r^2***	***Mean Absolute Error***
Temp_Comp Temp_Solvent	Mean Particle Size **(MPS)**	0.258	264.7
Temp_Comp Temp_Solvent hydophilic_PF		0.167	273.3
Compound_Type Temp_Comp Temp_Solvent		0.129	289.51
hydrophilic_sphericity hydrophobic_sphericity Full_mol_spherciity hydophilic_pf (individual or combined)	**Polydispersity**	-1.79	0.229
hydrophilic_sphericity hydrophobic_sphericity hydrophilic_packing	**Ratio Sphere Substance**	-2.001	0.289

Case (3): Prediction using **Synthesis Conditions + RD KIT**

Predictor	***Target***	***r^2***	***Mean Absolute Error***
Temp_Solvent Temp_compound	Mean Particle Size **(MPS)**	0.262	268.62
Temp_Solvent Temp_compound hydophilic_packingfact		0.174	277.03
PEOE_VSA6_S or TPSA_C Temp_comp temp_solvent		0.135	295.81
Estate_VSA8_S SlogP_VSA4_S hydrophilic_sphericity hydrophobic_sphericity NumAliphaticHeterocycles NoCount	**Polydispersity**	-1.790	0.226
VSA_EState2 **or** Estate_VSA8_C **or** SlogP_VSA4_C **or** Nocount **or** numalpiphaitcrings **or** numalpihiaitrings **or** TPSA **or**	**Ratio Sphere Substance**	-0.771	0.225
hydophilic_packingfact or hydrophilic_sphericity or		-1.982	0.295

Full mol _sphericity			

END

Excess Information

TSPA = topological polar surface area (Molecular polar surface area (PSA))

PEOE_VSA1
PEOE_VSA2 – 5 - 10
PEOE_VSA3 – 10 - -15
PEOE_VSA6 = MOE -

Brief Results

- To predict Mean Particle Size – RD Kit descriptors had least r2 score
- To predict Dispersity – Combined (Synthesis conditions + RD Kit Desc) was the ideal choice with best r2 score of 0.3
- Temperature is frequently used across
- Selective RD Kit decriptors are
- Compound also plays a crucial role
- We need to focus on working combinations and term the rest as unfit/junk
-

Discussions

- **Removal of AFM metrics**
- **Removal of Packing Factors**

1. Surface **Behaviour** of **Calixarene** under various **compositions** & Synthesis conditions
2. **Theoretical Descriptors** to **predict sphere properties**
3. **Models** to select most **relevant metrics/descriptors** that can **assess** the **characteristics** of occurring **sphere**
4. Analysis of large Calixarene dataset (20 compounds) for Spheres

---
