Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.


# Machine learning and molecular dynamics simulations assisted evolutionary design and discovery pipeline to screen the efficient small molecule acceptors for PTB7-Th-based organic solar cells with over 15% efficiency

Asif Mahmood[a], Ahmad Irfan[b] and Jin-Liang Wang*[a]

[a]Key Laboratory of Cluster Science of Ministry of Education, Beijing Key Laboratory of Photoelectronic/Electrophotonic Conversion Materials, School of Chemistry and Chemical Engineering, Beijing Institute of Technology, Beijing, 100081, China.

[b]Department of Chemistry, College of Science, King Khalid University, P.O. Box 9004, Abha 61413, Saudi Arabia

E-mail: jinlwang@bit.edu.cn

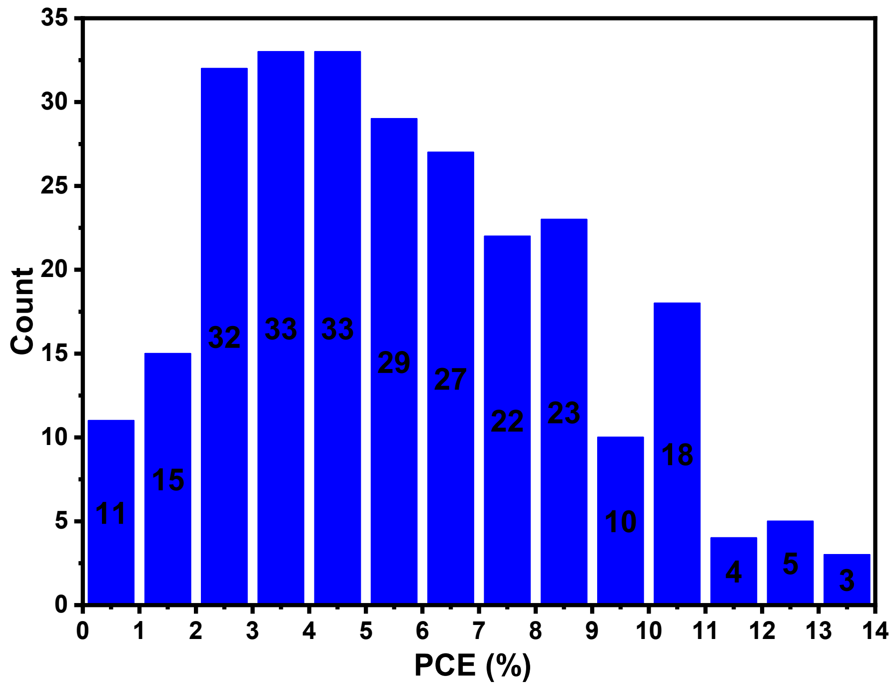



Figure S1. Distribution of experimental efficiencies of PBT7-Th-based OSCs in the data set.

## Major categories of descriptors

The descriptors are the languages that scientists use to communicate with machine learning models. Descriptors along the target property are used as input to trained machine models. Our studied materials are organic molecules. The molecular descriptors are suitable for machine leaning. Molecular descriptors are thousands in numbers. The major categories are given below:

Autocorrelation descriptors

Basak descriptors

BCUT descriptors

Burden descriptors

Connectivity descriptors

Constitutional descriptors

E-state descriptors

Kappa descriptors

Molecular property descriptors

Quantum chemical descriptors

Topological descriptors

CPSA descriptors

RDF descriptors

Geometrical descriptors

WHIM descriptors

Autocorrelation (2D and 3D) descriptors

Constitutional indices

Ring descriptors

matrix-based descriptors (2D and 3D)

MoRSE descriptors

GETAWAY descriptors

These categories also have sub categories.

## Baseline modelling

A baseline is a method that uses heuristics, simple summary statistics, randomness, or machine learning to create predictions for a dataset. Predictions from baseline modelling are used to measure the baseline performance and used as metric to compare the performance of other machine learning algorithms.

The baseline modelling is performed using Zero Rule algorithm in Weka. The results are given in Table S1 and S2. 3, 5, 10 and 20 fold cross-validation is used. Cross-validation has not shown strong effect on mean absolute error (MAE) and root mean square error (RMSE).

Table S1. Mean absolute error (MAE) of baseline model for various target properties.

| Target property | PCE (%) | HOMO (eV) | LUMO (eV) | $\lambda_{max(s)}$ (nm) | $\lambda_{max(f)}$ (nm) |
|---|---|---|---|---|---|
| 3-fold | 2.528 | 0.237 | 0.221 | 79 | 87 |
| 5-fold | 2.532 | 0.237 | 0.221 | 79 | 87 |
| 10-fold | 2.535 | 0.237 | 0.221 | 79 | 87 |
| 20-fold | 2.531 | 0.237 | 0.221 | 79 | 87 |

Table S2. Root mean square error (RMSE) of baseline model for various target properties.

| Target property | PCE (%) | HOMO (eV) | LUMO (eV) | $\lambda_{max(s)}$ (nm) | $\lambda_{max(f)}$ (nm) |
|---|---|---|---|---|---|
| 3-fold | 3.030 | 0.300 | 0.288 | 99 | 109 |
| 5-fold | 3.036 | 0.300 | 0.288 | 99 | 109 |
| 10-fold | 3.039 | 0.300 | 0.288 | 99 | 109 |
| 20-fold | 3.036 | 0.300 | 0.288 | 99 | 109 |

Table S3. Descriptors and their correlation between them and PCE.

| No. | Name | Category | Description | Correlation with PCE |
|---|---|---|---|---|
| 1 | ATSC6s | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 6 weighted by I-state | +0.576 |
| 2 | MATS6s | 2D autocorrelations | Moran autocorrelation of lag 6 weighted by I-state | +0.544 |
| 3 | RDF040m | RDF descriptor | Radial distribution function-040/ weighted by relative mass | +0.540 |
| 4 | G(O..F) | 3D Atom Pairs | sum of geometrical distances between O..F | +0.521 |
| 5 | MDEC-34 | Topological descriptors | Molecular distance edge between all tertiary and quaternary carbons | +0.505 |
| 6 | VCH-6 | Connectivity descriptors | Valence chain, order 6 | +0.501 |

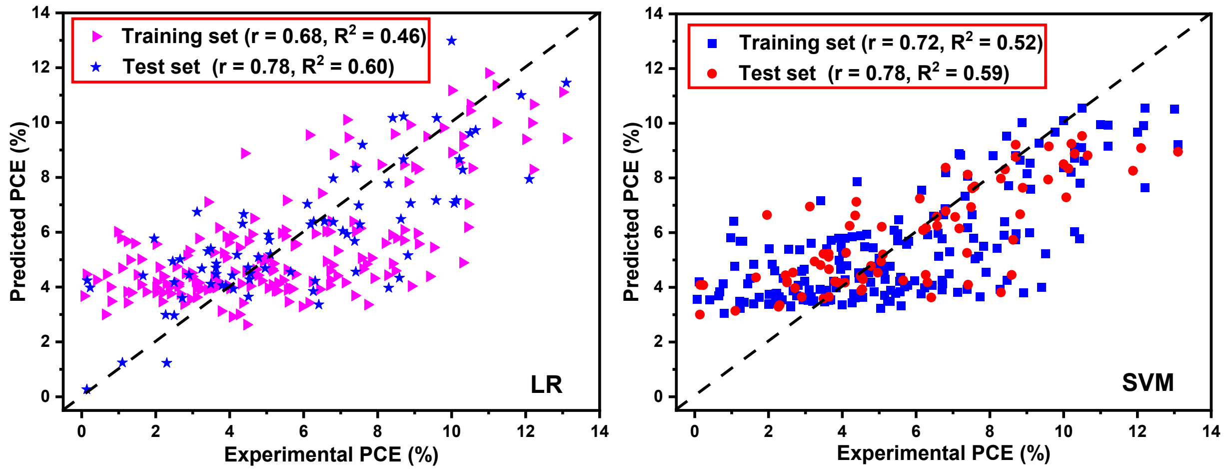



Figure S2. Pearson correlation between experimental PCE and predicted PCE calculated using linear regression (LR) and support vector machine (SVM)

Table S4. Root mean square error (RMSE), mean absolute error (MAE), coefficient of determination ($R^2$) and Pearson correlation coefficient (r) of various machine learning models for the prediction of PCE.

| Model | | RMSE (%) | MAE (%) | $R^2$ | r |
|---|---|---|---|---|---|
| k-nearest neighbor | Training set* | 1.93 | 1.54 | 0.60 | 0.77 |
| | Test set | 1.97 | 1.56 | 0.57 | 0.76 |
| Linear Regression | Training set* | 2.22 | 1.81 | 0.46 | 0.68 |
| | Test set | 1.88 | 1.47 | 0.60 | 0.78 |
| Random Forest | Training set* | 1.19 | 0.91 | 0.84 | 0.93 |
| | Test set | 1.12 | 0.87 | 0.86 | 0.93 |
| Support vector machine | Training set* | 2.10 | 1.68 | 0.52 | 0.72 |
| | Test set | 1.92 | 1.47 | 0.59 | 0.78 |

*Cross-validation = (3-fold)

Table S5. Descriptors and their correlation between them and HOMO.

| No. | Name | Category | Description | Correlation with HOMO |
|---|---|---|---|---|
| 1 | SM1_Dzs | Topological descriptors | Spectral moment of order 1 from Barysz matrix / weighted by I-state | +0.502 |
| 2 | ETA_dAlpha_B | Topological descriptors | A measure of count of hydrogen bond acceptor atoms and/or | +0.474 |

| | | | | |
|---|---|---|---|---|
| | | | polar surface area | |
| 3 | RCI | Ring descriptors | ring complexity index | +0.471 |
| 4 | J_D/Dt | 2D matrix-based descriptors | Balaban-like index from distance/detour matrix | +0.434 |
| 5 | SpMAD_EA(ri) | Edge adjacency indices | spectral mean absolute deviation from edge adjacency mat. weighted by resonance integral | +0.426 |
| 6 | Mor04v | 3D-MoRSE descriptors | signal 04 / weighted by van der Waals volume | +0.425 |
| 7 | SpMAD_AEA(dm) | Edge adjacency indices | spectral mean absolute deviation from augmented edge adjacency mat. weighted by dipole moment | +0.417 |

Table S6. Descriptors and their correlation between them and LUMO.

| No. | Name | Category | Description | Correlation with LUMO |
|---|---|---|---|---|
| 1 | SpMax5_Bh(s) | Burden eigenvalues | largest eigenvalue n. 5 of Burden matrix weighted by I-state | +0.491 |
| 2 | SpMAD_AEA(dm) | Edge adjacency indices | spectral mean absolute deviation from augmented edge adjacency mat. weighted by dipole moment | +0.473 |
| 3 | Eig07_AEA(dm) | Edge adjacency indices | eigenvalue n. 7 from augmented edge adjacency mat. weighted by dipole momen | +0.472 |
| 4 | R1s | GETAWAY descriptors | R autocorrelation of lag 1 / weighted by I-state | +0.463 |
| 5 | SM5_B(s) | 2D matrix-based descriptors | spectral moment of order 5 from Burden matrix weighted by I-State | +0.462 |
| 6 | AVS_B(s) | 2D matrix-based descriptors | average vertex sum from Burden matrix weighted by I-State | +0.457 |
| 7 | P_VSA_s_6 | P_VSA-like descriptors | P_VSA-like on I-state, bin 6 | +0.449 |
| 8 | RPSA | CPSADescriptor | TPSA / total molecular surface area | +0.443 |
| 9 | P_VSA_LogP_5 | P_VSA-like descriptors | P_VSA-like on LogP, bin 5 | +0.441 |

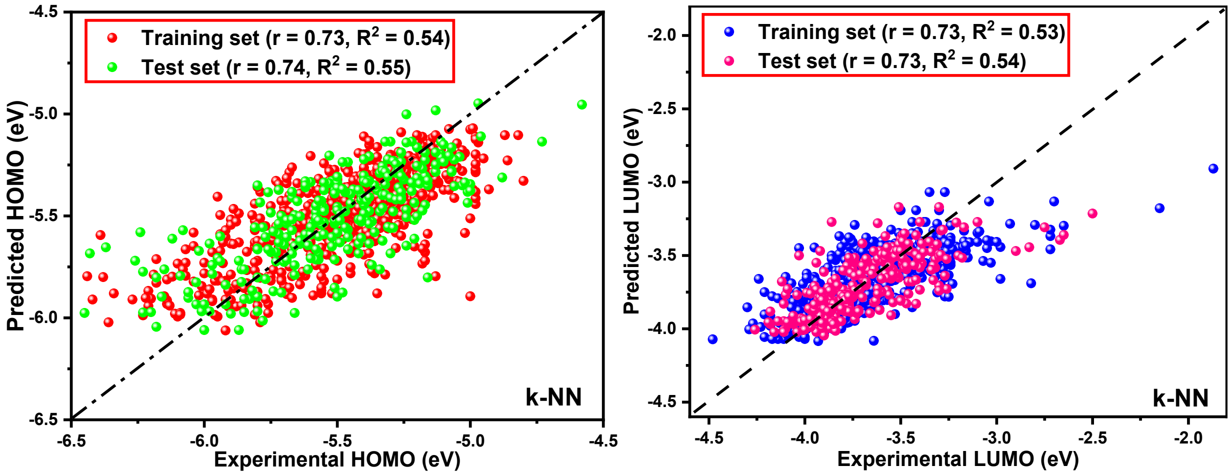



Figure S3. Correlation between experimental and predicted energy levels using k-nearest neighbor (k-NN)

Table S7. Root mean square error (RMSE), mean absolute error (MAE), coefficient of determination ($R^2$) and Pearson correlation coefficient (r) of k-nearest neighbor and Random Forest models for the prediction of HOMO and LUMO.

| Energy level | Model | | RMSE (eV) | MAE (eV) | $R^2$ | r |
|---|---|---|---|---|---|---|
| HOMO | k-nearest neighbor | Training set* | 0.20 | 0.15 | 0.54 | 0.73 |
| | | Test set | 0.21 | 0.16 | 0.55 | 0.74 |
| | Random Forest | Training set* | 0.11 | 0.08 | 0.86 | 0.93 |
| | | Test set | 0.12 | 0.09 | 0.84 | 0.92 |
| LUMO | k-nearest neighbor | Training set* | 0.20 | 0.15 | 0.53 | 0.73 |
| | | Test set | 0.19 | 0.14 | 0.54 | 0.73 |
| | Random Forest | Training set* | 0.11 | 0.08 | 0.85 | 0.93 |
| | | Test set | 0.11 | 0.07 | 0.85 | 0.93 |

*Cross-validation = (10-fold)

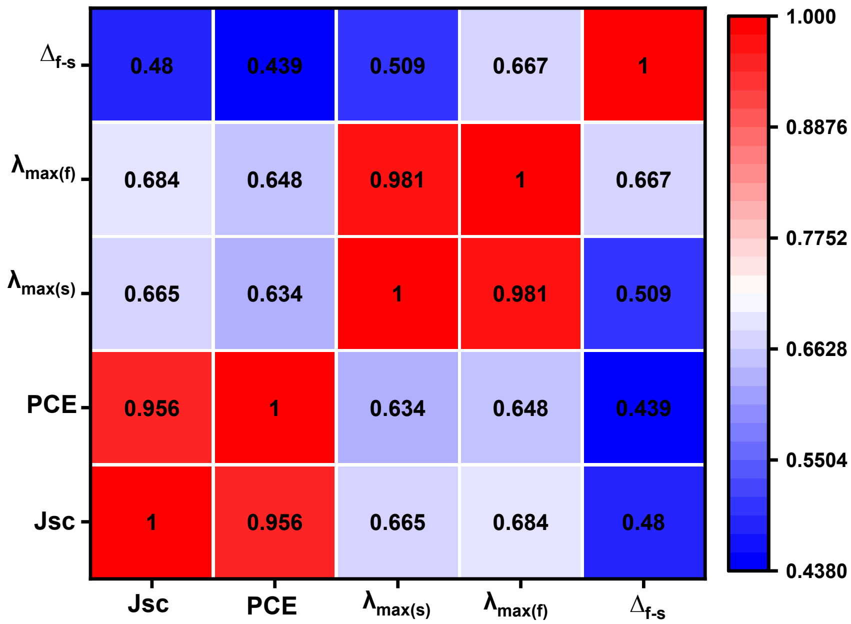



Figure S4. Heatmap of correlation between various parameters of OSCs: short circuit current (Jsc), power conversion efficiency (PCE), absorption maxima of UV/visible spectra determined in solution ($\lambda_{max(s)}$), absorption maxima of UV/visible spectra determined in solution ($\lambda_{max(f)}$) and difference between $\lambda_{max(s)}$ and $\lambda_{max(f)}$ ($\Delta_{f\text{-}s}$)

Table S8. Descriptors and their correlation between them and $\lambda_{max(s)}$.

| No. | Name | Category | Description | Correlation with $\lambda_{max(s)}$ |
|---|---|---|---|---|
| 1 | nR08 | Ring descriptors | number of 8-membered rings | +0.626 |
| 2 | SpDiam_AEA(ri) | Edge adjacency indices | spectral diameter from augmented edge adjacency mat. weighted by resonance integral | +0.579 |
| 4 | C-017 | Atom-centred fragments | =CR2 | +0.510 |
| 5 | D/Dtr08 | Ring descriptors | distance/detour ring index of order 8 | +0.496 |
| 6 | ATSC5s | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 5 weighted by I-state | +0.494 |
| 7 | MDEC-34 | Topological descriptors | Molecular distance edge between all tertiary and quaternary carbons | +0.481 |
| 8 | SpMax_A | 2D matrix-based descriptors | leading eigenvalue from adjacency matrix (Lovasz- | +0.479 |

| | | | | |
|---|---|---|---|---|
| | | | Pelikan index) | |
| 9 | JGI3 | 2D autocorrelations | mean topological charge index of order 3 | +0.462 |
| 10 | ATSC6s | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 6 weighted by I-state | +0.460 |

Table S9. Descriptors and their correlation between them and $\lambda_{max(f)}$.

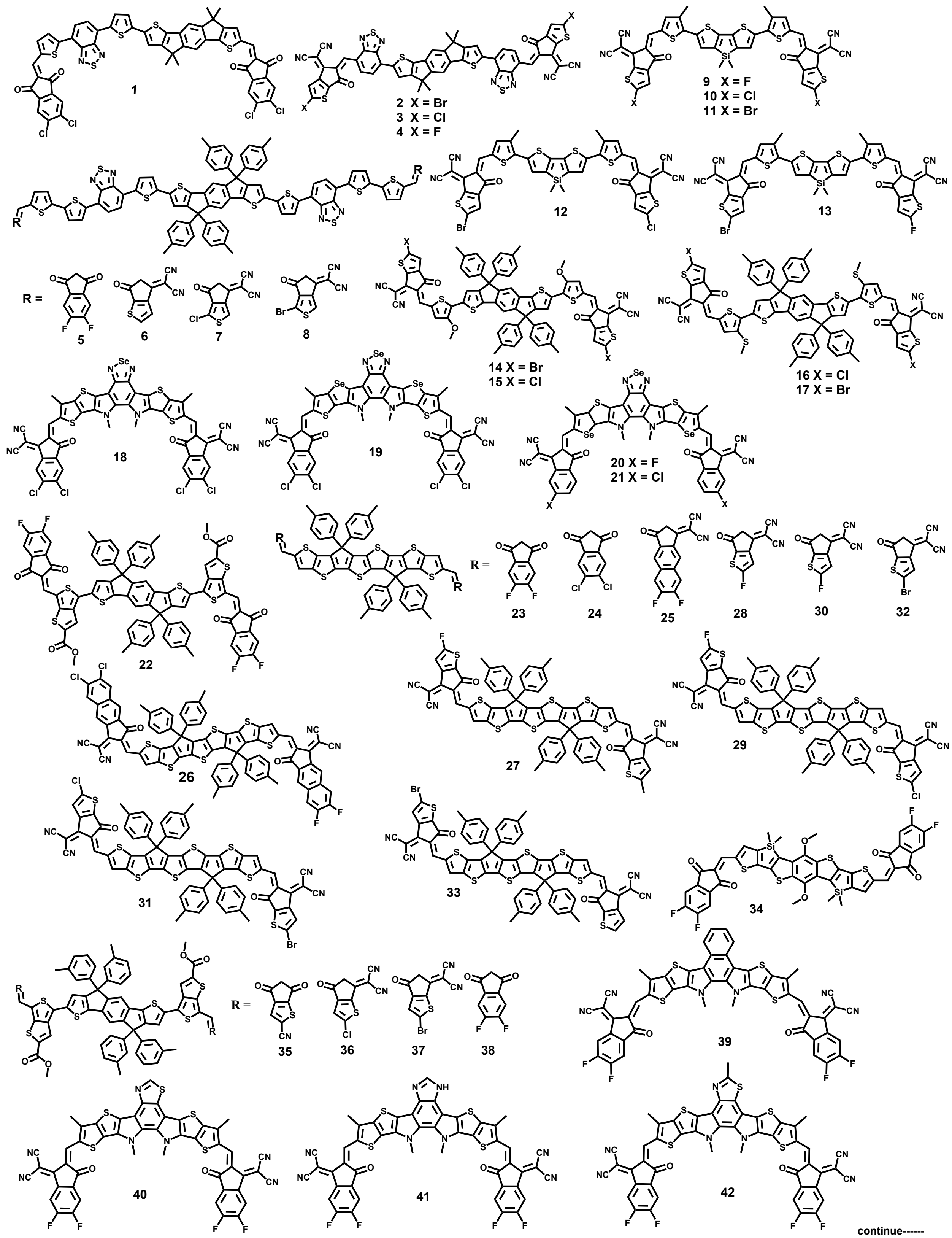
| No. | Name | Category | Description | Correlation with $\lambda_{max(f)}$ |
|---|---|---|---|---|
| 1 | nR08 | Ring descriptors | number of 8-membered rings | +0.586 |
| 2 | SpDiam_AEA(ri) | Edge adjacency indices | spectral diameter from augmented edge adjacency mat. weighted by resonance integral | +0.548 |
| 3 | C-017 | Atom-centred fragments | =CR2 | +0.501 |
| 4 | IC1 | Information indices | Information Content index (neighborhood symmetry of 1-order) | +0.479 |
| 5 | CATS2D_06_AA | CATS 2D | CATS2D Acceptor-Acceptor at lag 06 | +0.471 |
| 6 | JGI3 | 2D autocorrelations | mean topological charge index of order 3 | +0.460 |
| 7 | ATSC5s | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 5 weighted by I-state | +0.458 |
| 8 | ATSC6s | 2D autocorrelations | Centred Broto-Moreau autocorrelation of lag 6 weighted by I-state | +0.435 |

Table S10. Root mean square error (RMSE), mean absolute error (MAE), coefficient of determination ($R^2$) and Pearson correlation coefficient (r) of k-nearest neighbor and Random Forest models for the prediction of absorption maxima in solution and film.

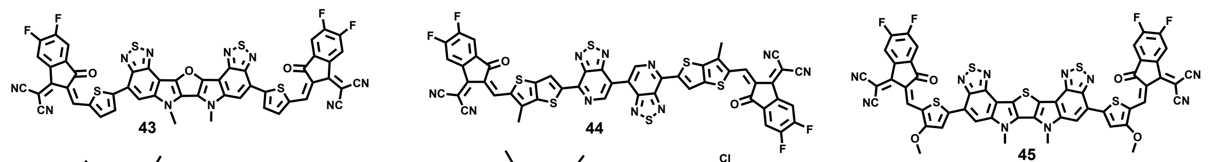
| | Model | | RMSE (nm) | MAE (nm) | $R^2$ | r |
|---|---|---|---|---|---|---|
| solution | k-nearest neighbor | Training set* | 55.48 | 40.11 | 0.68 | 0.83 |
| | | Test set | 57.1 | 42.14 | 0.68 | 0.83 |
| | Random Forest | Training set* | 27.88 | 18.78 | 0.92 | 0.96 |
| | | Test set | 33.33 | 22.33 | 0.90 | 0.96 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Film | k-nearest neighbor | Training set* | 73.99 | 56.85 | 0.53 | 0.73 |
| | | Test set | 67.12 | 52.16 | 0.64 | 0.83 |
| | Random Forest | Training set* | 32.90 | 24.01 | 0.91 | 0.96 |
| | | Test set | 35.55 | 25.78 | 0.90 | 0.96 |

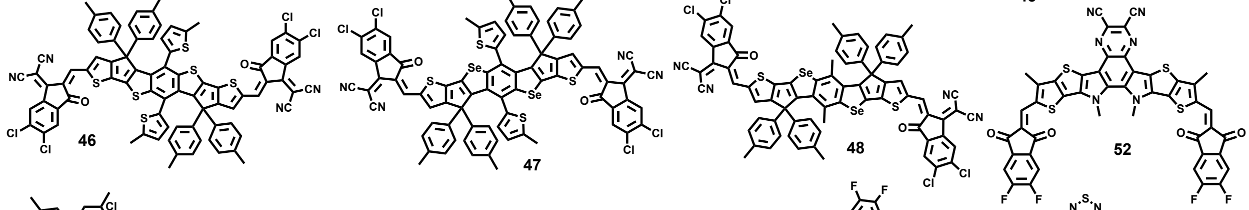
*Cross-validation = (3-fold)


1
2 X = Br
3 X = Cl
4 X = F
9 X = F
10 X = Cl
11 X = Br
12
13
R =
5
6
7
8
14 X = Br
15 X = Cl
16 X = Cl
17 X = Br
18
19
20 X = F
21 X = Cl
22
R =
23
24
25
28
30
32
26
27
29
31
33
34
R =
35
36
37
38
39
40
41
42

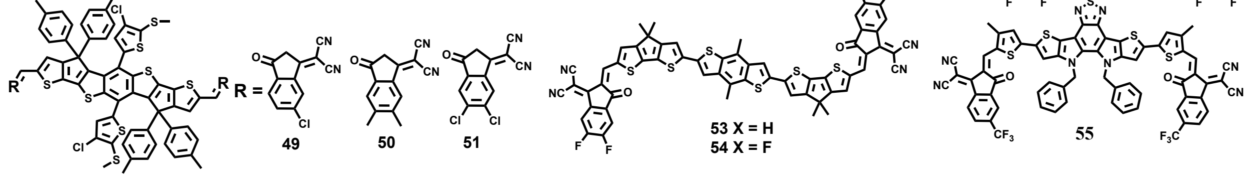
continue------



43 44 45



46 47 48 52

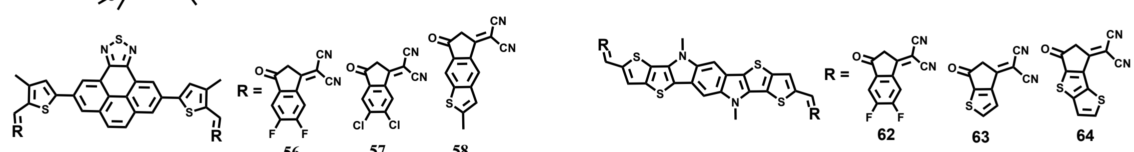


R =

49 50 51

53 X = H
54 X = F

55

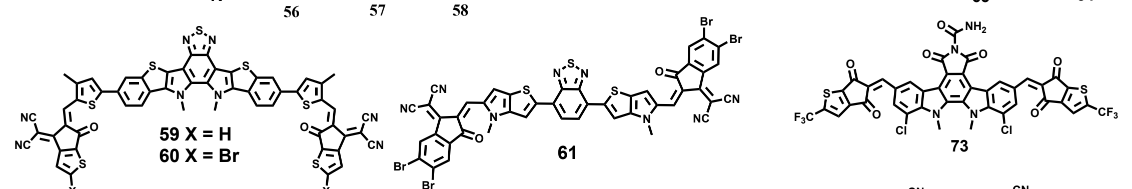


R =

56 57 58

R =

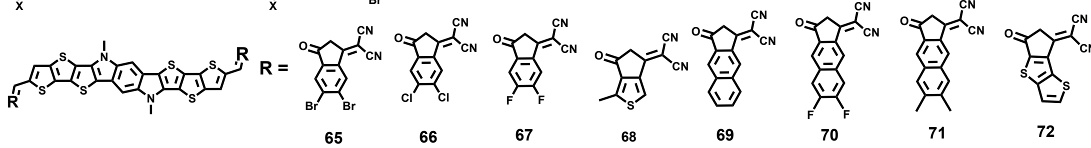
62 63 64



59 X = H
60 X = Br

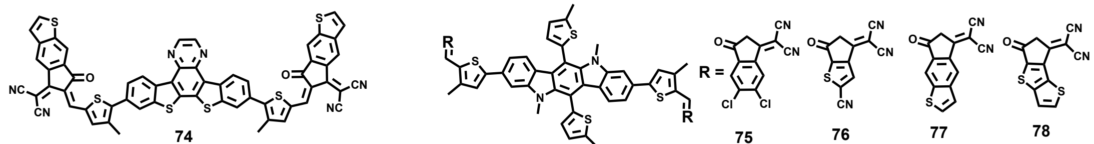
61

73



R =

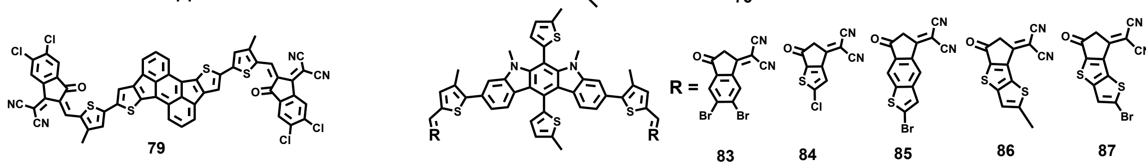
65 66 67 68 69 70 71 72



74

R =

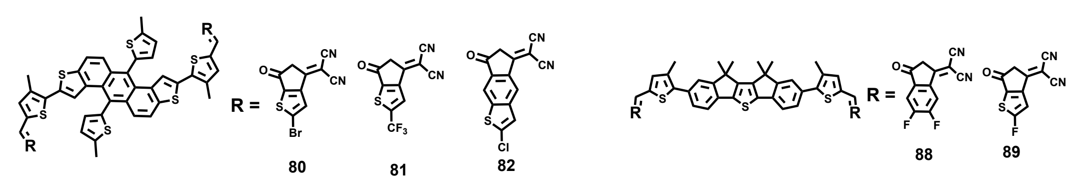
75 76 77 78



79

R =

83 84 85 86 87



R =

80 81 82

R =

88 89

continue....

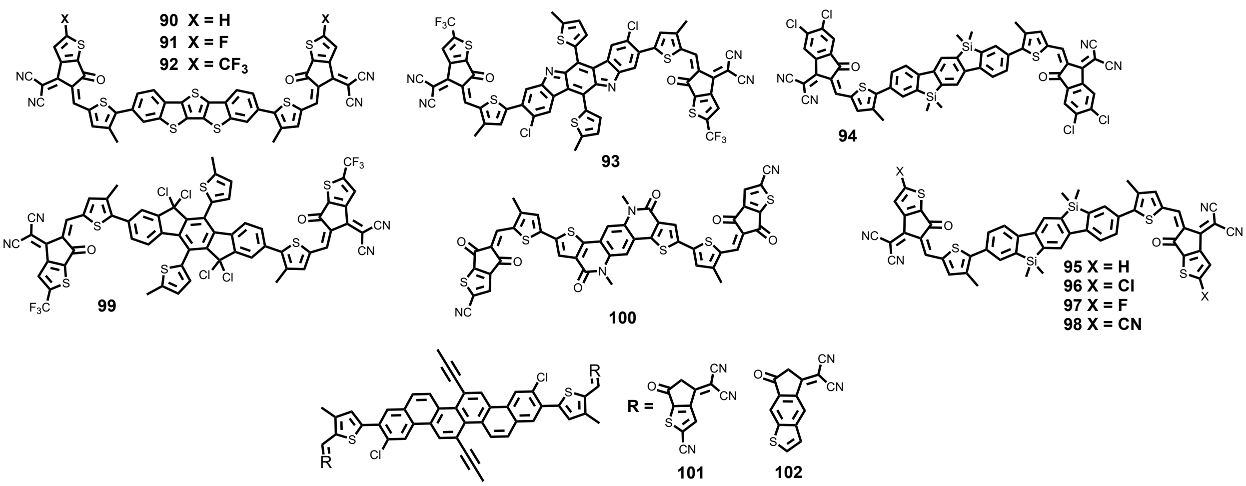



Figure S5. Potential candidates of new SMAs for PBT7-Th: SMAs based OSCs.

Table S11. Energy level, absorption maxima, PCE and Florgy-Huggins parameter ($\chi$) for more than 100 potential candidates.

| No. | HOMO (eV) | LUMO (eV) | $\lambda_{max(s)}$ (nm) | $\lambda_{max(f)}$ (nm) | PCE (%) | $\chi$ |
|---|---|---|---|---|---|---|
| 1 | -5.52 | -3.86 | 653 | 731 | 13.17 | -31.81 |
| 2 | -5.45 | -3.78 | 698 | 793 | 14.57 | 42.40 |
| 3 | -5.43 | -3.82 | 698 | 793 | 14.24 | -10.01 |
| 4 | -5.44 | -3.89 | 698 | 793 | 13.66 | 3.20 |
| 5 | -5.26 | -3.59 | 692 | 654 | 14.57 | -9.01 |
| 6 | -5.22 | -3.67 | 671 | 749 | 13.01 | -27.03 |
| 7 | -5.25 | -3.56 | 671 | 737 | 13.78 | 5.30 |
| 8 | -5.24 | -3.68 | 671 | 740 | 14.75 | 37.01 |
| 9 | -5.44 | -3.85 | 725 | 724 | 13.81 | 23.07 |
| 10 | -5.46 | -3.90 | 737 | 765 | 14.27 | 23.40 |
| 11 | -5.46 | -3.89 | 732 | 765 | 13.80 | 10.07 |
| 12 | -5.46 | -3.89 | 732 | 777 | 14.27 | 13.13 |
| 13 | -5.46 | -3.82 | 725 | 748 | 14.09 | 32.24 |
| 14 | -5.39 | -3.94 | 701 | 739 | 14.52 | 19.97 |
| 15 | -5.40 | -3.97 | 701 | 739 | 14.23 | 14.17 |
| 16 | -5.53 | -3.86 | 701 | 739 | 14.94 | 8.918 |
| 17 | -5.58 | -3.76 | 694 | 748 | 15.26 | 38.07 |
| 18 | -5.53 | -3.86 | 725 | 775 | 13.94 | 32.60 |
| 19 | -5.54 | -3.96 | 716 | 775 | 13.94 | -8.84 |
| 20 | -5.55 | -4.01 | 733 | 777 | 13.50 | -14.11 |
| 21 | -5.56 | -3.97 | 731 | 778 | 13.64 | -12.38 |
| 22 | -5.49 | -3.65 | 673 | 646 | 14.32 | -7.63 |
| 23 | -5.45 | -3.66 | 643 | 653 | 15.51 | 26.58 |
| 24 | -5.36 | -3.69 | 629 | 733 | 13.17 | 23.09 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 25 | -5.51 | -3.99 | 790 | 789 | 13.09 | 3.98 |
| 26 | -5.48 | -3.98 | 790 | 795 | 13.09 | 23.43 |
| 27 | -5.54 | -3.95 | 790 | 785 | 13.03 | 13.17 |
| 28 | -5.53 | -4.02 | 790 | 785 | 13.30 | 23.00 |
| 29 | -5.58 | -4.02 | 790 | 785 | 13.13 | 40.18 |
| 30 | -5.56 | -3.91 | 790 | 776 | 13.28 | 23.74 |
| 31 | -5.53 | -3.91 | 787 | 776 | 13.28 | 24.56 |
| 32 | -5.57 | -3.92 | 785 | 776 | 13.28 | 9.48 |
| 33 | -5.52 | -3.90 | 785 | 776 | 13.28 | 16.97 |
| 34 | -5.54 | -3.76 | 672 | 675 | 13.81 | 6.63 |
| 35 | -5.37 | -3.71 | 667 | 771 | 13.35 | 14.36 |
| 36 | -5.38 | -4.04 | 679 | 775 | 14.80 | 35.65 |
| 37 | -5.38 | -3.84 | 670 | 775 | 14.80 | 25.28 |
| 38 | -5.42 | -3.57 | 690 | 634 | 14.96 | 15.99 |
| 39 | -5.67 | -3.95 | 740 | 795 | 13.52 | 22.23 |
| 40 | -5.54 | -3.99 | 732 | 795 | 13.52 | 29.42 |
| 41 | -5.82 | -3.90 | 732 | 795 | 13.66 | 19.17 |
| 42 | -5.54 | -3.97 | 757 | 801 | 13.52 | -13.87 |
| 43 | -5.60 | -3.91 | 727 | 777 | 14.07 | 16.87 |
| 44 | -5.60 | -3.85 | 577 | 777 | 13.52 | -11.26 |
| 45 | -5.73 | -3.87 | 727 | 768 | 13.10 | 26.31 |
| 46 | -5.39 | -3.94 | 751 | 740 | 15.24 | 3.99 |
| 47 | -5.43 | -3.88 | 745 | 729 | 15.24 | 24.91 |
| 48 | -5.48 | -4.02 | 722 | 723 | 13.43 | -3.29 |
| 49 | -5.49 | -3.83 | 754 | 739 | 15.49 | -0.12 |
| 50 | -5.53 | -3.74 | 756 | 751 | 12.86 | 12.81 |
| 51 | -5.48 | -3.83 | 756 | 736 | 15.53 | 51.60 |
| 52 | -5.56 | -4.00 | 723 | 679 | 13.52 | 19.35 |
| 53 | -5.45 | -3.84 | 701 | 753 | 13.92 | 75.84 |
| 54 | -5.51 | -3.90 | 702 | 759 | 13.60 | 53.24 |
| 55 | -5.64 | -3.91 | 659 | 760 | 13.83 | 28.50 |
| 56 | -5.73 | -3.86 | 585 | 605 | 14.25 | 5.09 |
| 57 | -5.53 | -3.84 | 579 | 592 | 13.70 | -1.39 |
| 58 | -5.61 | -3.83 | 592 | 585 | 13.13 | 70.46 |
| 59 | -5.52 | -3.82 | 716 | 823 | 13.69 | 23.15 |
| 60 | -5.48 | -3.81 | 707 | 807 | 14.67 | 34.57 |
| 61 | -5.67 | -3.72 | 562 | 775 | 13.82 | -37.54 |
| 62 | -5.62 | -3.83 | 733 | 793 | 13.48 | 24.24 |
| 63 | -5.42 | -3.94 | 731 | 791 | 13.70 | 62.8 |
| 64 | -5.44 | -3.98 | 730 | 797 | 14.09 | -4.34 |
| 65 | -5.42 | -3.86 | 712 | 801 | 14.76 | 40.65 |
| 66 | -5.37 | -3.88 | 709 | 773 | 14.76 | -2.30 |
| 67 | -5.46 | -3.93 | 716 | 802 | 14.25 | 107.50 |
| 68 | -5.48 | -3.96 | 721 | 780 | 14.61 | 18.34 |
| 69 | -5.54 | -3.81 | 702 | 757 | 13.74 | 44.74 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 70 | -5.51 | -3.89 | 716 | 807 | 13.47 | 32.21 |
| 71 | -5.56 | -3.81 | 701 | 745 | 13.86 | -7.75 |
| 72 | -5.46 | -3.91 | 703 | 794 | 14.30 | 69.20 |
| 73 | -5.43 | -3.94 | 612 | 753 | 13.00 | 5.79 |
| 74 | -5.46 | -3.84 | 594 | 611 | 13.43 | 111.10 |
| 75 | -5.45 | -3.90 | 605 | 609 | 13.46 | 16.88 |
| 76 | -5.42 | -3.78 | 655 | 725 | 13.05 | -3.71 |
| 77 | -5.43 | -3.91 | 591 | 599 | 13.75 | 28.17 |
| 78 | -5.52 | -3.82 | 719 | 766 | 13.10 | 4.58 |
| 79 | -5.47 | -3.80 | 717 | 723 | 13.5 | -42.57 |
| 80 | -5.40 | -3.79 | 702 | 767 | 14.66 | 42.03 |
| 81 | -5.46 | -3.95 | 692 | 740 | 13.41 | -30.81 |
| 82 | -5.44 | -3.73 | 578 | 623 | 13.25 | -11.35 |
| 83 | -5.48 | -3.89 | 605 | 616 | 13.68 | 13.29 |
| 84 | -5.46 | -3.79 | 724 | 724 | 14.10 | -13.75 |
| 85 | -5.48 | -3.93 | 568 | 639 | 14.40 | 15.50 |
| 86 | -5.54 | -3.82 | 726 | 795 | 13.22 | 31.29 |
| 87 | -5.50 | -3.82 | 726 | 803 | 13.72 | 9.38 |
| 88 | -5.53 | -3.87 | 711 | 754 | 13.81 | -5.07 |
| 89 | -5.40 | -3.83 | 699 | 722 | 13.76 | 21.49 |
| 90 | -5.47 | -3.82 | 569 | 751 | 13.33 | 65.85 |
| 91 | -5.37 | -3.79 | 583 | 715 | 13.95 | 42.58 |
| 92 | -5.35 | -3.84 | 594 | 775 | 13.89 | -29.00 |
| 93 | -5.48 | -3.88 | 694 | 793 | 13.17 | -9.53 |
| 94 | -5.48 | -3.88 | 604 | 640 | 13.11 | 59.89 |
| 95 | -5.53 | -3.83 | 727 | 732 | 13.32 | 32.62 |
| 96 | -5.52 | -3.80 | 725 | 725 | 14.25 | 4.46 |
| 97 | -5.51 | -3.95 | 730 | 719 | 14.25 | 28.47 |
| 98 | -5.51 | -3.91 | 730 | 712 | 13.32 | 45.17 |
| 99 | -5.44 | -3.86 | 684 | 743 | 14.47 | -41.56 |
| 100 | -5.37 | -3.89 | 502 | 695 | 13.63 | 41.92 |
| 101 | -5.60 | -3.89 | 572 | 725 | 13.51 | 111.70 |
| 102 | -5.34 | -3.80 | 487 | 598 | 14.06 | 101.40 |