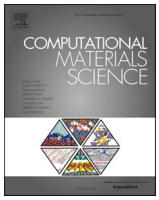
Computational Materials Science 241 (2024) 113037

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci



Full Length Article

Machine learning-assisted chemical space generation of small molecule organic semiconductors for efficient photodetectors

Khadijah Mohammedsaleh Katubi [a], Alvi Muhammad Rouf [b,*], Bilal Siddique [b], Muhammad Faizan Nazar [b,*], Ghulam Jillani Ansari [c], Z.A. Alrowaili [d], M.S. Al-Buriahi [e]

[a] Department of Chemistry, College of Science, Princess Nourah bint Abdulrahman University, P.O. Box: 84428, Riyadh 11671, Saudi Arabia
[b] Department of Chemistry, Division of Science and Technology, University of Education, Lahore, 54770, Pakistan
[c] Department of Information Sciences, Division of Science and Technology, University of Education, Lahore, 54770, Pakistan
[d] Department of Physics, College of Science, Jouf University, P.O. Box: 2014, Sakaka, Saudi Arabia
[e] Department of Physics, Sakarya University, Sakarya, Turkey

ARTICLE INFO

Keywords:
Machine learning
Chemical similarity
Photodetectors
Semiconductors
Sensors

ABSTRACT

Organic photodetectors (OPDs), the emerging next-generation candidates for light sensing applications, could compensate deficiencies of the traditional inorganic photodetectors. To get OPDs of highly improved performance, and broader applications, machine learning (ML) technique is employed. A chemical space of 20,000 organic semiconductors (OSCs) by RDKit *via* BRICKS (Breaking Retrosynthetically Interesting Chemical Substructures) method was achieved, from the selected building blocks of the OSC data obtained from the literature, contributing to big data. Out of these, top twenty OSCs were screened based on synthetic accessibility score, chemical similarity analysis, and structural behavior understanding.

1. Introduction

Photodetectors (PDs), also defined as photosensors, mainly function through converting energy of UV–vis, IR, and other electromagnetic radiations into electrical signals, and measure properties of light through photoelectric effect [1,2]. These photoelectric conversion techniques in the form of sensors or detectors have remarkable applications in the fields of optical communications, image registration, environmental monitoring, and biological and chemical testing [3,4]. However, the most commercial PDs use photodiodes, which are composites of inorganic semiconductors, such as, silicon or germinum compounds (due to their high charge-carrier mobility, high stability, and low exciton binding energy) [5]. However, these also have their own disadvantages, for instance, expensive and complicated manufacturing processes, and mechanical inflexibility, that severely limits practical applications of the commercial inorganic PDs. Therefore, organic PDs (OPDs), the emerging next-generation candidates for light sensing, compensate deficiencies of traditional inorganic PDs due to high substrate flexibility, ease of processing, light weight, tunable absorption properties, and low-cost manufacturing process [6,7]. Fundamentally, organic semiconductors (OSCs) are soft-structured materials with larger band gaps, discrete energy levels, and better absorptivity due to weaker van der Waals interactions between assembled molecules, making them highly suitable candidates for the fabrication of optoelectronic devices (e.g., solar cells and OPDs). The OSCs are made up of repeating units of π-bonded molecules or polymers composed of hydrocarbons, and sometimes heteroatoms, such as, oxygen, nitrogen, and sulpher [8,9]. The OSCs differ in composition and structure from traditional inorganic semiconductors due to the impact of bonding forces between atoms/molecules, which result in significant differences in their respective mechanical, electrical and optical properties [10,11]. Nowadays, performance capabilities of OPDs are constantly getting improved to widen their applicabilions.

The traditional material design is usually based on trial and error approach. Synthesis and characterization continues until the required results are achived. However, it is an extremely lethargic strategy that, sometimes, requires lengthy experimental procedure(s), lot of materials and complex analytical techniques, making the whole scenario extremely expensive, and time-consuming [12,13]. To overcome these challenges, computer-aided material discovery, designing, and screening are becoming famous [14]. Computational methods use mathematical algorithms to analyze and predict scientific data. They can

* Corresponding authors.
E-mail addresses: rouf.alvi@ue.edu.pk (A. Muhammad Rouf), faizan.nazar@ue.edu.pk (M. Faizan Nazar).

https://doi.org/10.1016/j.commatsci.2024.113037
Received 10 January 2024; Received in revised form 18 April 2024; Accepted 18 April 2024
Available online 25 April 2024

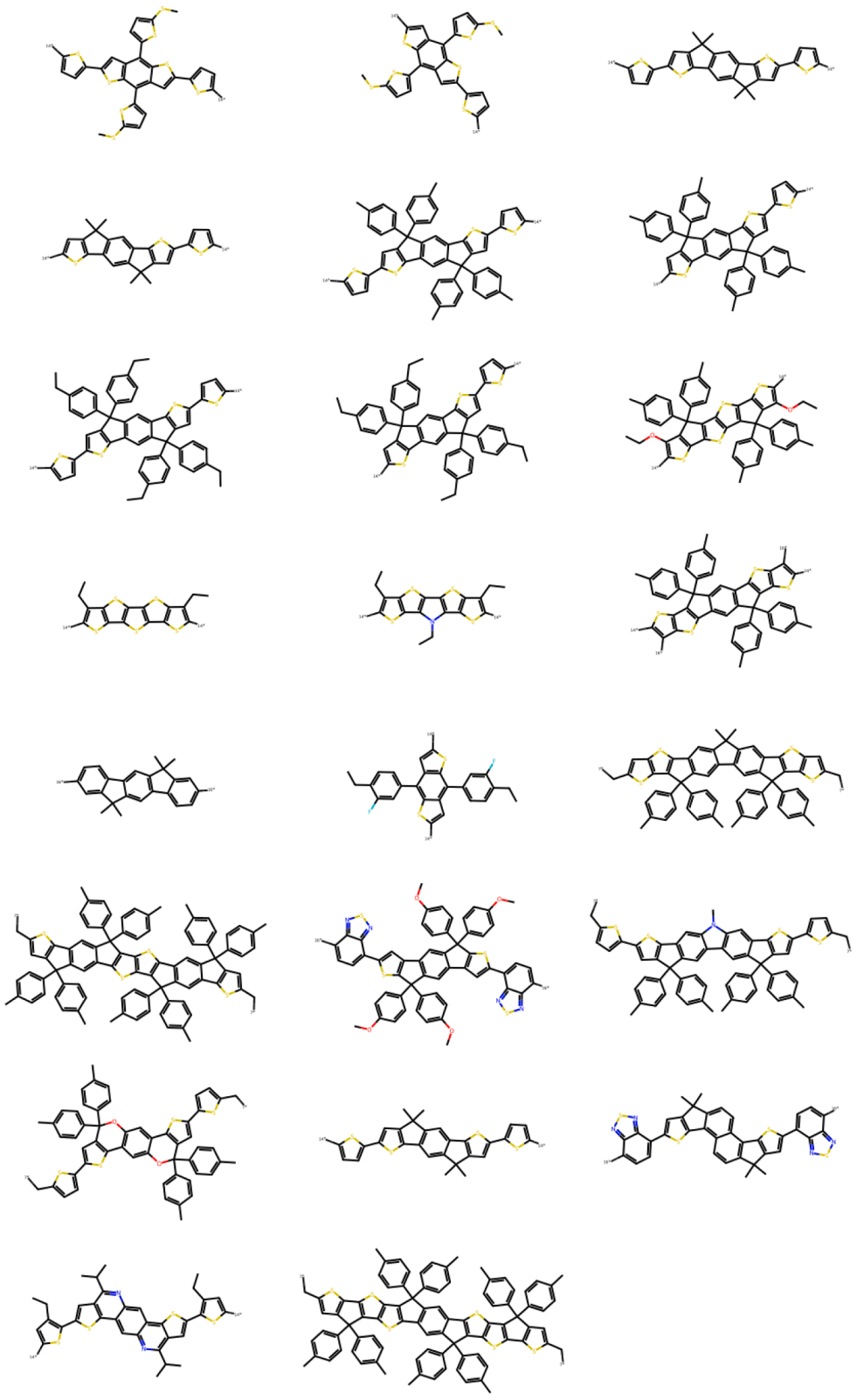


Fig. 1. Central building blocks used to design OSCs.

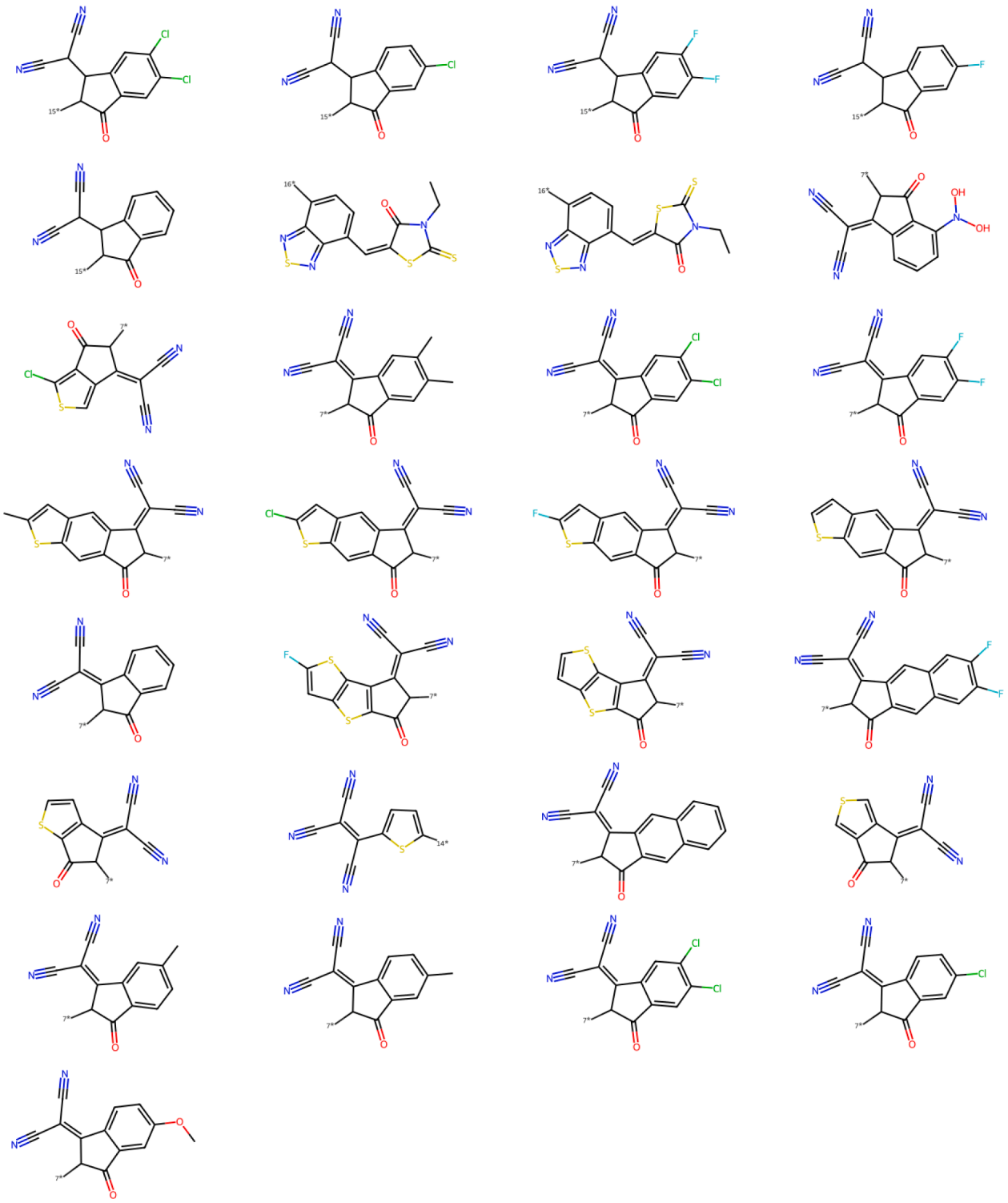


Fig. 2. Terminal building blocks used to design OSCs.

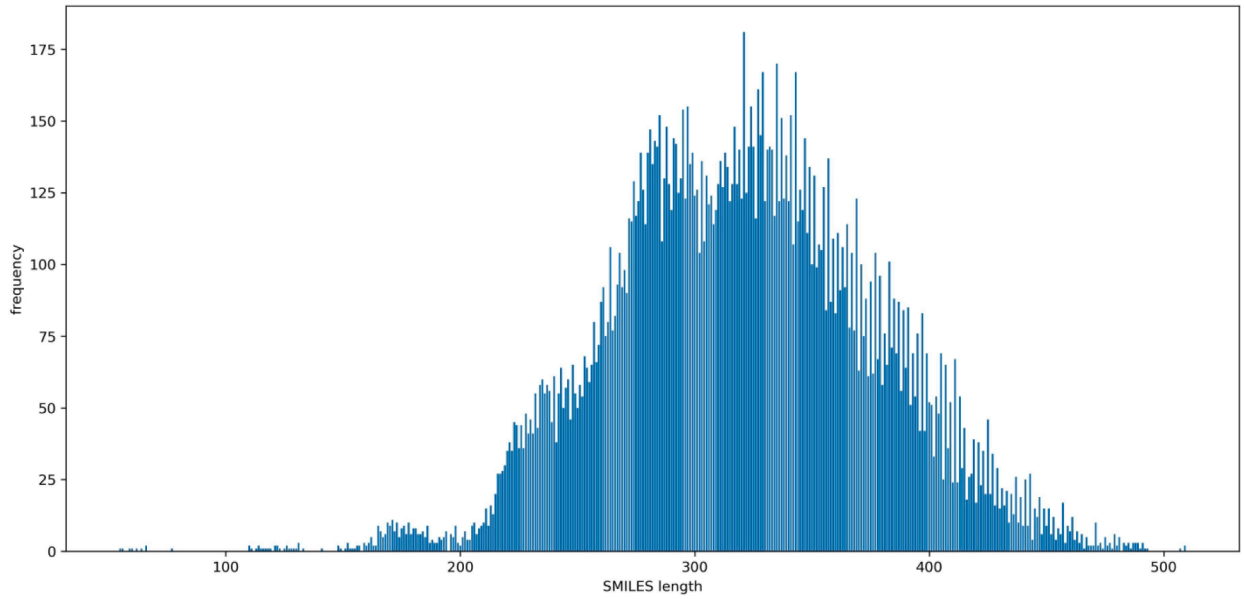


Fig. 3. SMILES length distruntion and frequency.

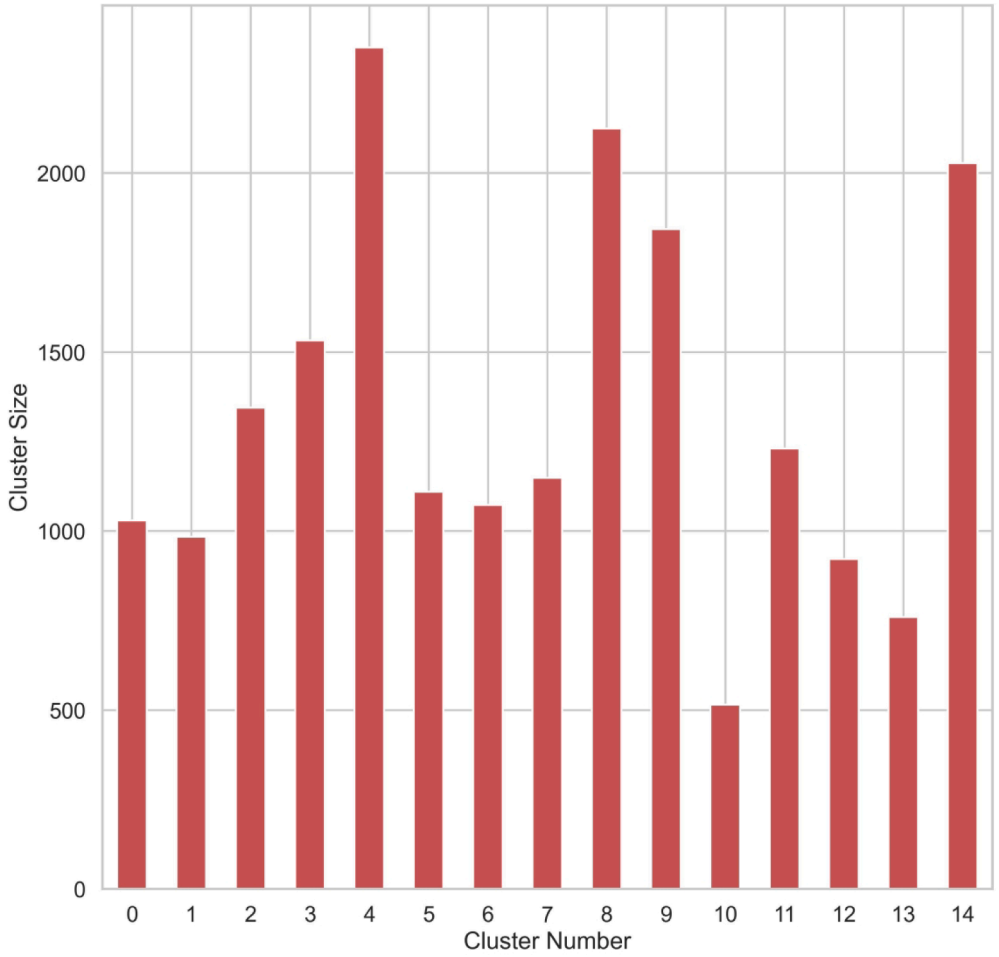


Fig. 4. Distribution of monomers in different clusters.

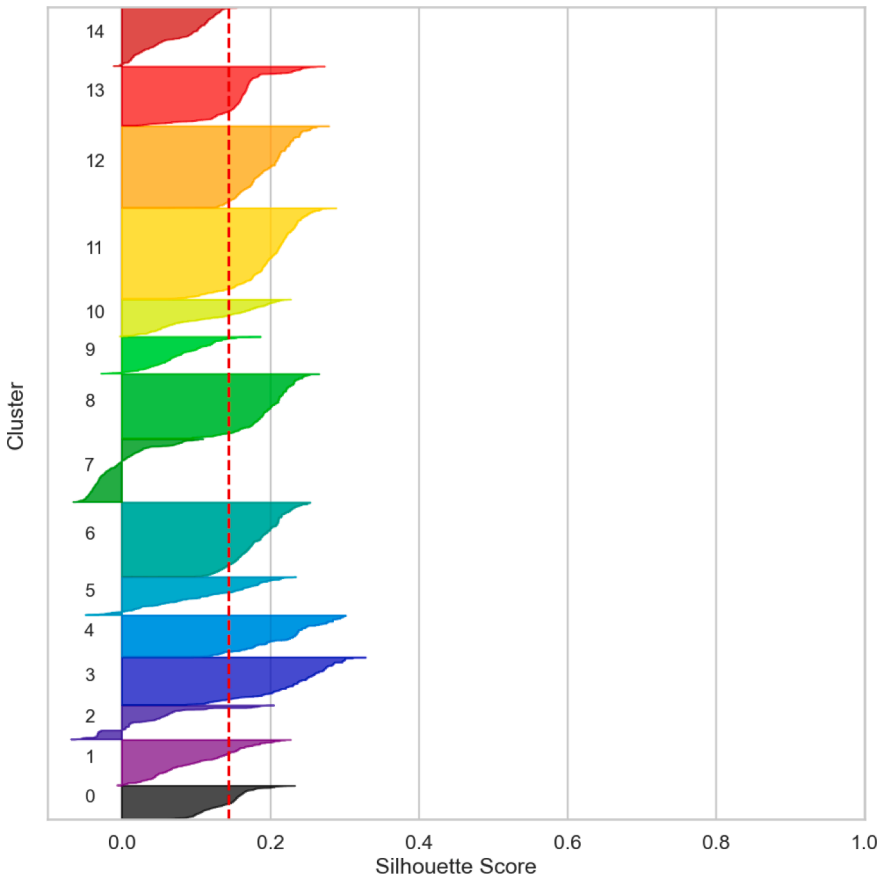


Fig. 5. Silhouette plot of clusters.

efficiently solve complicated scientific problems of various systems [15,16]. Over the past decade, researchers have focused on developing the predictive computer models to investigate and resolve challenging tasks [17]. A step further, artificial intelligence (AI) is now being dubbed as the "fourth model of science" [18]. Machine learning (ML) is a modern research tool and has been at the core of AI since the 1980 s due to its ability to reorganize existing knowledge of structures and discover implicit relationships between them. For material science, ML has emerged as a powerful tool to design and screen valuable compounds for superconductors, photovoltaics, and high entropy alloys [19]. Subsequently, these molecules were successfully subjected to experimental testing. ML also facilitated and accelerated development of efficient materials for OSCs, thus improving their properties in a short-time.

Conjugated polymers and small molecules both are used as the active materials in OPDs. An OSC contains p-type donor which is intermixed with n-type acceptor. The charge generation upon photo excitation in an OPD is based on the heterojunction layer between the donor–acceptor molecules, and internal photoelectric effect defines an OPD's working principle. The OSC absorbs energy of incoming light photons and converts into electrical current. Both polymer- and π-conjugated molecular-based OSCs display high values of extinction coefficients due to the large wave function – an overlap between electronic ground state and lowest excited state. It allows OSCs to absorb light strongly in the form of flexible and thin films. The absorption of photon causes the generation of excitons, which are diffused to the interface between the donor and acceptor to produce free charges.

In the current study, a chemical space of new OSCs was generated by automatic method. Synthetic accessibility of generated OSCs was calculated. OSCs are screened on the basis of synthetic accessibility score. Chemical smiliarity analysis is performed on slected OSCs.

2. Computational details

We designed unpredented OSCs employing RDKit [20] *via* Breaking Retrosynthetically Interesting Chemical Substructures (BRICS) approach from 20,000 semiconductors by using selected building blocks. BRICS was used to break compounds into fragments, and then these fragments were joined together in various combinations to propose new ones. This approach does systematic fragmentation under predefined guidelines carefully. RDKit is also employed to calculated synthetic accessibility (SA) score. Further, chemical similarity analysis (CSA), and structural behavior understading is achieved by clustering and heat map, respectively. For that purpose, method of extended connectivity fingerprints of diameter 4 (ECFP4) was considered [21].

3. Results and discussion

3.1. Designing OSCs

The common, and easily synthesizable, electron-deficient and electron-rich groups are selected to design OSCs [22,23]. The chemical structures of the central buiding blocks are given in Fig. 1. These fused-ring building blocks are good to achieve higher charge transfer. Terminal building blocks bearing common electron-deficient groups are given in Fig. 2. As decribed above, BRICS method [24] has produced 20,000 OSCs using selected building blocks.

SMILES, simplified molecular-input line-entry system, is basically a way of describing chemical structures in the form of strings [25,26]. Complexity and size of a molecule determines the length of SMILES string. Simple molecule has small length of SMILES string, while the complex molecule has long string. The histogram like distribution is obtained when a graph is plotted between frequency of molecules *vs* the

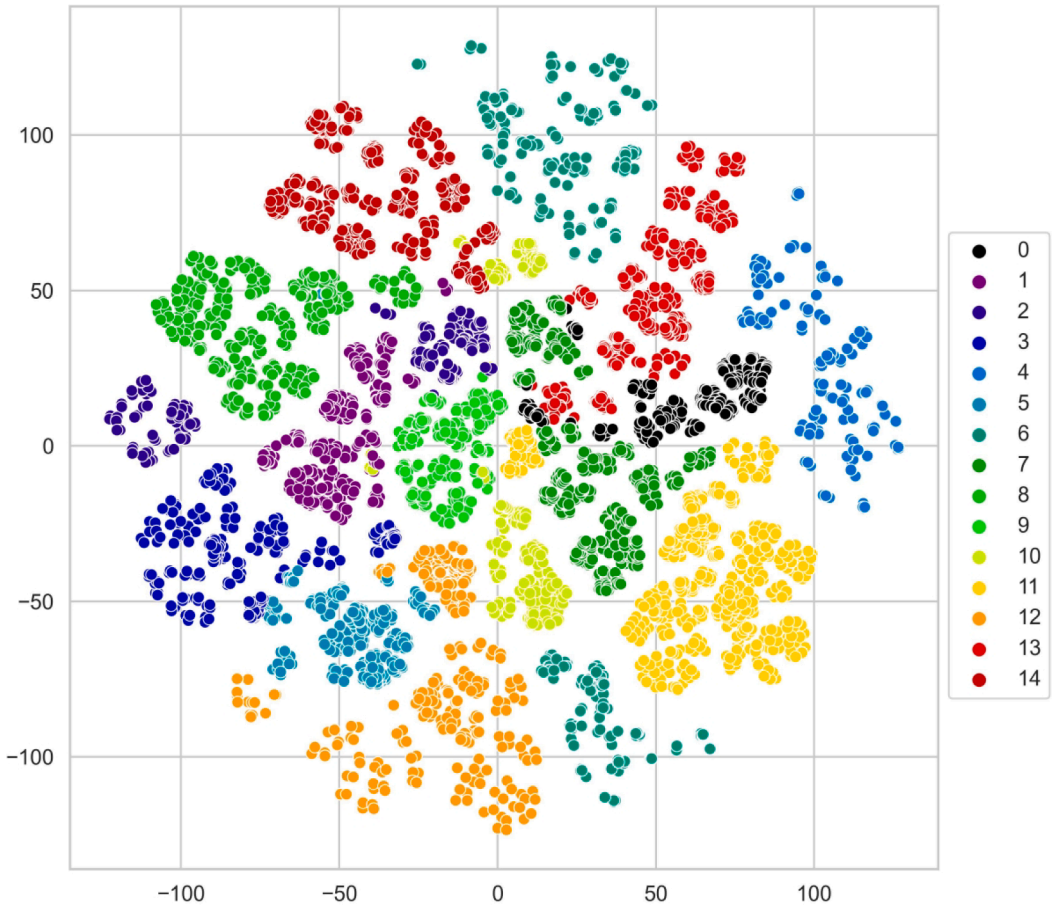


Fig. 6. The t-SNE plot of clusters.

Table 1
Synthetic accessibility (SA) score of the selected OSCs.

OSC	Synthetic accessibility score
1	1
2	1
3	1
4	1
5	1
6	1
7	1.14
8	1.94
9	2.14
10	2.28
11	2.53
12	2.54
13	2.56
14	2.67
15	2.78
16	2.8
17	2.83
18	2.84
19	2.87
20	2.89

SMILES length (Fig. 3). The graph has a bell curve shape, which means that most of the molecules have a SMILES length close to the average (250–400 characters), specifically, around 300 characters. The frequency of SMILES length decreases on moving away from the average, i.e., the lowest frequency around 25, and the highest frequency close to 175, indicating that the dataset has a normal distribution of SMILES length, with no outliers or skewness.

3.2. Cluster analysis of designed monomers

Clustering is a ML technique that identifies data in patterns, by grouping similar data points together, and keeping different data points apart. The relationship between cluster number and cluster size can be represented through a graph, typically a bar chart or a histogram. The graph in Fig. 4 shows size of 15 different clusters of data points. Each bar in the graph represents a cluster, and its height shows availability of data points in that cluster. The graph demonstrates some clusters are larger in size, indicating more data points there. For example, cluster number 4 has the largest size, with around more than 2000 data points. On the other hand, cluster number 10, a smaller sized cluster, carries around 500 data points.

The silhouette score evaluates an object's cohesion with its own cluster compared to the other clusters. Its value ranges from −1 to 1, indicating how well each point is separated from the other cluster, and matched to its own cluster. Well-cluster points gain high or positive silhouette score, and poor-cluster points or the point belonging to some other cluster show negative values. The clusters overlap or disperse, and cohesive or distinct is based on average silhouette coefficient, representing low and high value, respectively. The graph in Fig. 5 shows that cluster number 3 have very high silhouette score (around 0.4), and cluster number 14 have the lowest silhouette score (0.2). The latter cluster indicates less separation and cohesion, and also considered as lower quality cluster.

The t-distributed stochastic neighbor embedding (t-SNE) graph helps to understand the similarity and diversity of data points. It can also isolate potential groups or clusters of data points which share common characteristics. Fig. 6 displays distribution of 15 different classes (i.e., 0–14) of data points in a two-dimensional space. Each class is

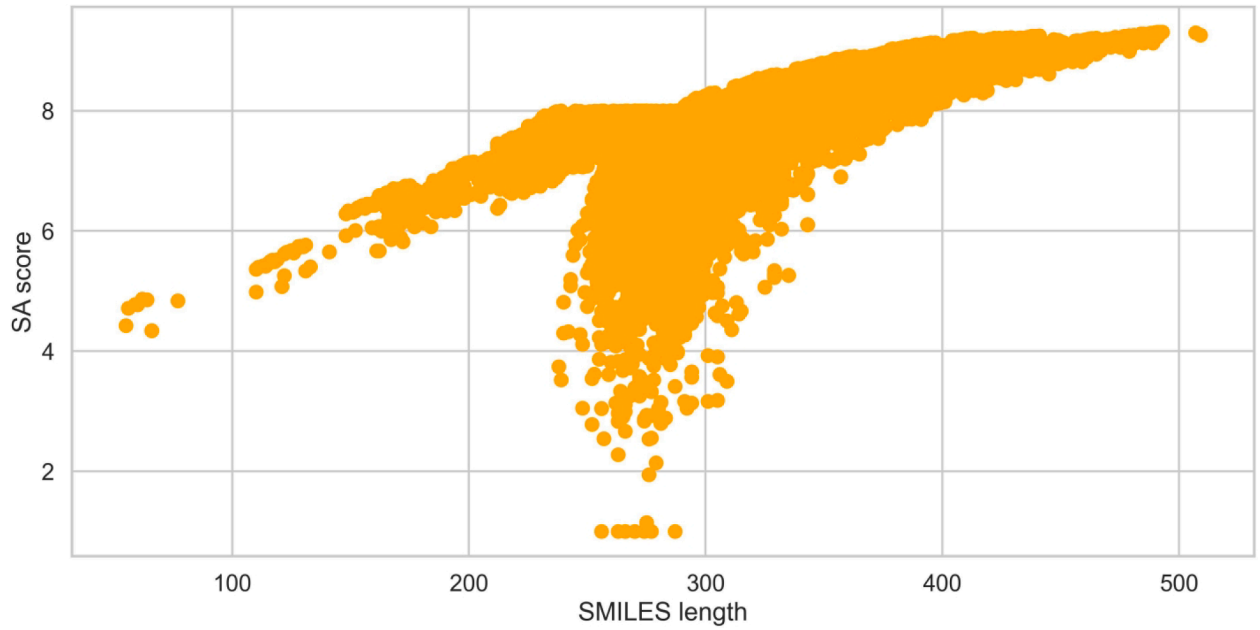


Fig. 7. Scatter plot between SA score and SMILES length.

represented by a different colour, and legend showing numbers assigned to the each class. In Fig. 6, some classes are much different from the others, thus, have less overlaping characteristics. Purple-colored class (i.e., 1) in the graph is mostly separated from the rest of the data. On the other hand, few classes are more scattered and mixed, which indicate more overlaping between them. Green-colored class (i.e., 7) is spread across the plot and interacts with many other classes. In the bottom left corner of the plot, a cluster formed by green- and orange-colored classes (i.e., 5 and 12, respectively) suggest common characteristics, however, they are distinct from the other classes.

3.3. Synthetic accessibility (SA)

The complexity of a compound is a key factor explaining its SA. For instance, a molecule bearing several functional groups is difficult to synthesize [27,28]. Furthermore, synthetic materials, methods, yields, purity, handling, and stability of reactants, intermediates, and products are all the vital factors determining SA. With the passage of time, several modern and innovative processes and strategies are devised to synthesized difficult molecules. The molecular modelling and combinatorial synthetic techniques can check the feasibility and physical structure of complex compounds before synthesis. The compounds which have low SA scores are easy to synthesize, while those heaving high SA scores indicate difficulty in their synthesis. Interestingly, Table 1 shows significantly low SA score (1.0 – 2.89) for the all twenty compounds, indicating their ease of synthesis. However, the scatter plot in Fig. 7 between SA score and SMILES length indicates no correlationship.

3.4. Screening the OSCs and similarity analysis

The synthesis of OSCs is influenced by the raw material that determines its properties, performance, quality, and cost [29,30]. Therefore, availability and affordability of the raw material is an important factor for the production of PDs, i.e., a cheap and easily accessible raw material can lower the production cost of PDs, leading to more economical and commercialized product, thus, enhancing supply, getting more revenue for innovation and development in the field [31,32]. We have designed approximately 20,000 OSCs using building block obtained from OSCs taken from the literature. Out of these, only twenty OSCs (shown in Figs. 8 and 9) were selected based on SA score, that could potentially be used in the production of OPDs.

Further, the similarity analysis was performed to compare similarity in chemical compounds through cheminformatics software, RDkit, based on their functional and structural connections [33,34]. Therefore, Fig. 10 indicates the clustering OSCs, where OSCs with similar information were found connected to each other displaying good similarity indices, and vice versa.

Additionally, a heatmap color-coding scheme is utilized to assess the relative similarity among the designed OSCs. Through the utilization of ML models, heatmaps visualize the evolving relationships between the characteristics of a model, enabling straightforward evaluation of the association between two variables. Fig. 11 illustrates the color pattern of the heatmap, depicting the distribution of OSCs based on their similarity. The scale ranges from lower blue colors, indicating a similarity value close to 0.82, to higher yellow colors with values nearing 1. OSCs with a similarity value close to 1, represented by the yellow color, indicate a higher degree of similarity, with most of the designed OSCs falling within the green region of the heatmap (value $>$ 0.94), underscoring their high similarity. Such similarity analysis proves valuable in both screening and designing potential candidates for OSC fabrication.

4. Conclusions

Organic photodetectors (OPDs), or organic photosensors are the modern developing candidates for light sensing due to high substrate flexibility, light weight, processing ease, tunable absorption properties, and low-cost manufacturing process. To get OPDs of highly improved performance, and broader applications, machine learning (ML) approch is employed. Conjugated polymers and small molecules both are used as the active materials in OPDs. However, the common and easy to synthesize electron-deficient and electron-rich groups are selected to design OSCs. These fused-ring building blocks are good to achieve higher charge transfer. RDKit *via* BRICS method has produced unprecedented 20,000 OSCs using selected building blocks, creating a new chemical space that is a contribution to big data. Heatmap and cluster plot are used to peformed the similarity analysis. Interestingly, calculated synthetic accessibility score highlighted considerable ease in the synthesis of the predicted OSCs. Our curent ML-based study leads to chemical space generation of OSCs.

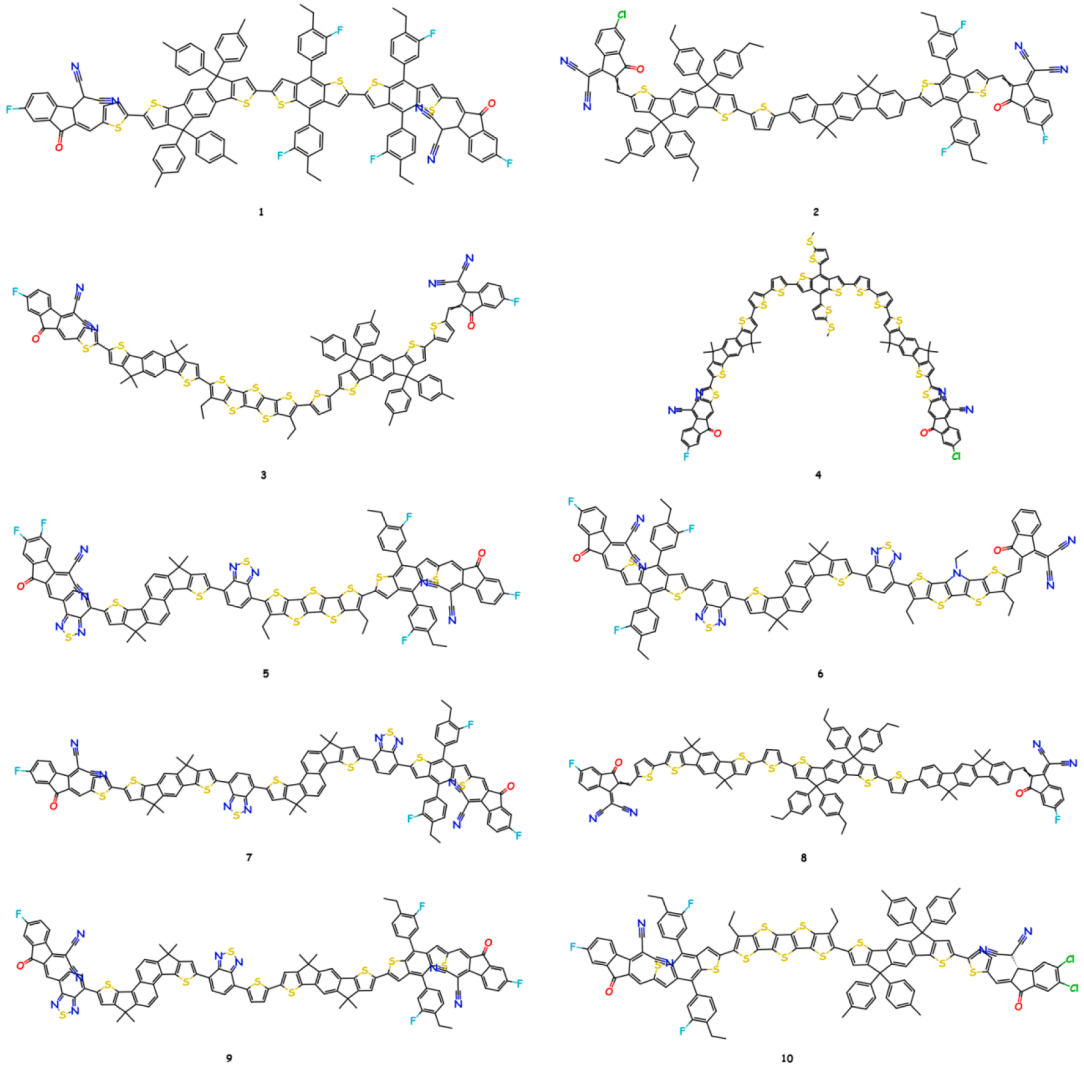


Fig. 8. Selected OSCs **1**–**10**.

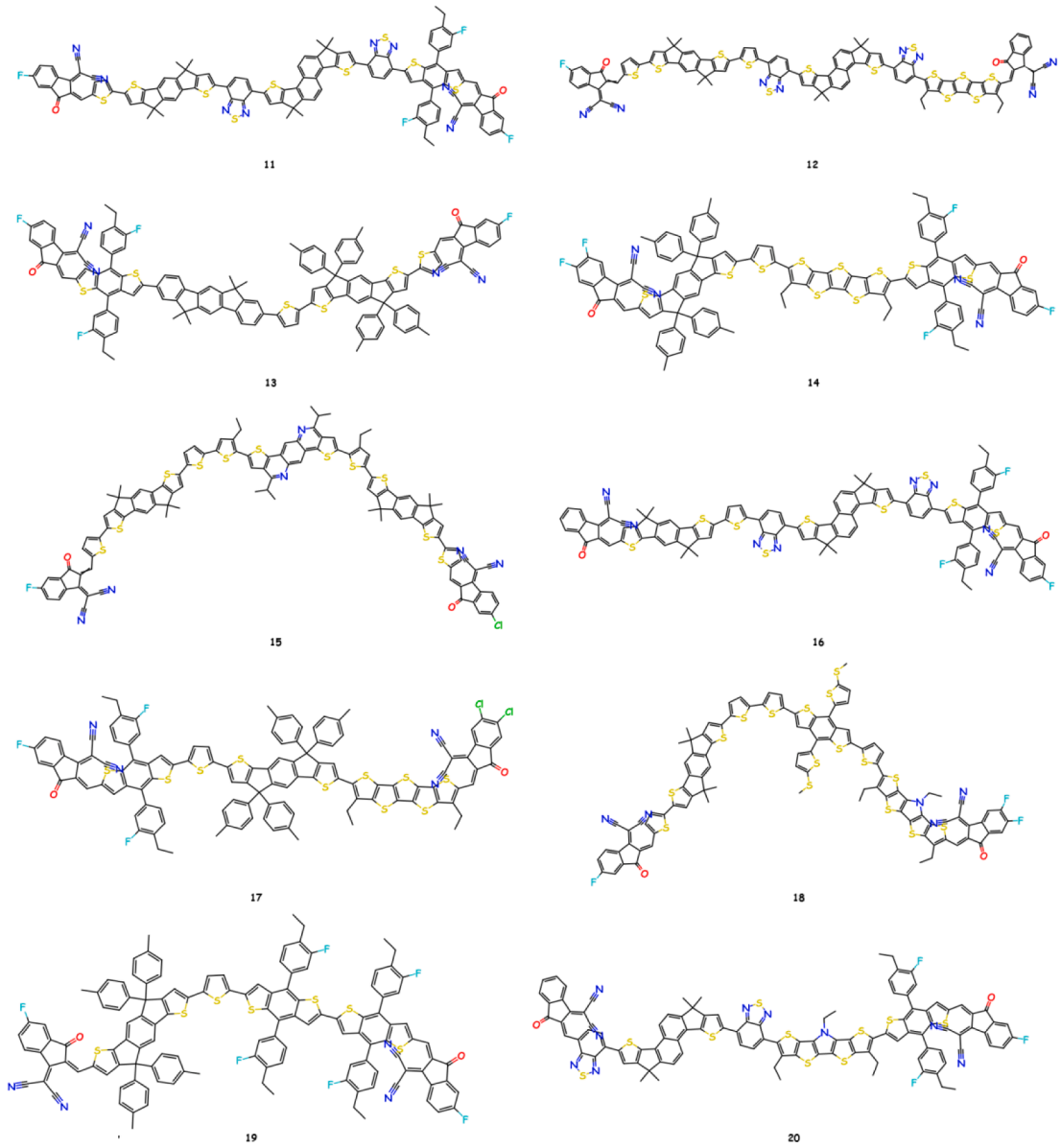


Fig. 9. Selected OSCs **11**–**20**.

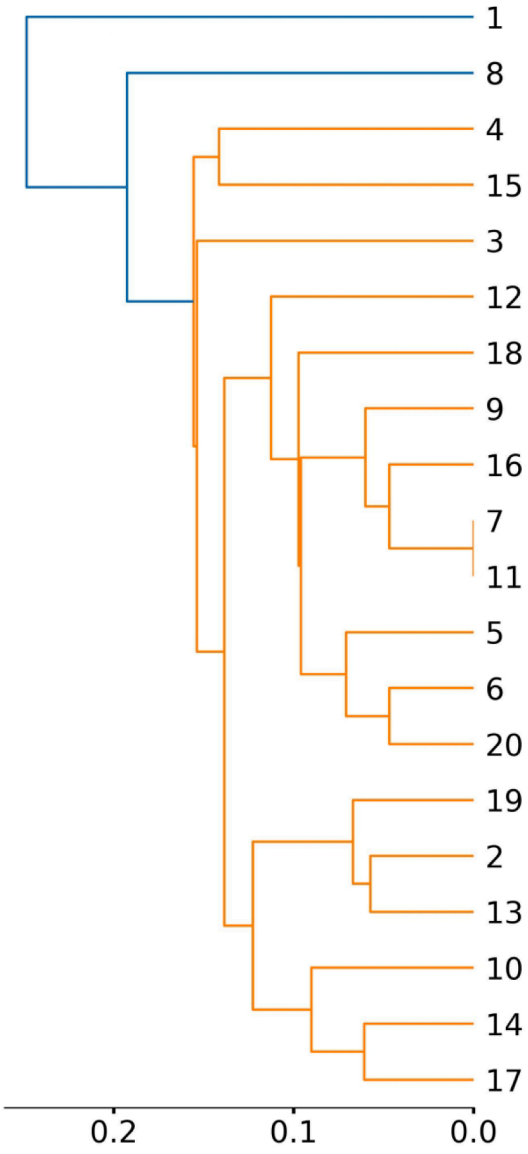


Fig. 10. Clustering of OSCs based on similarity analysis.

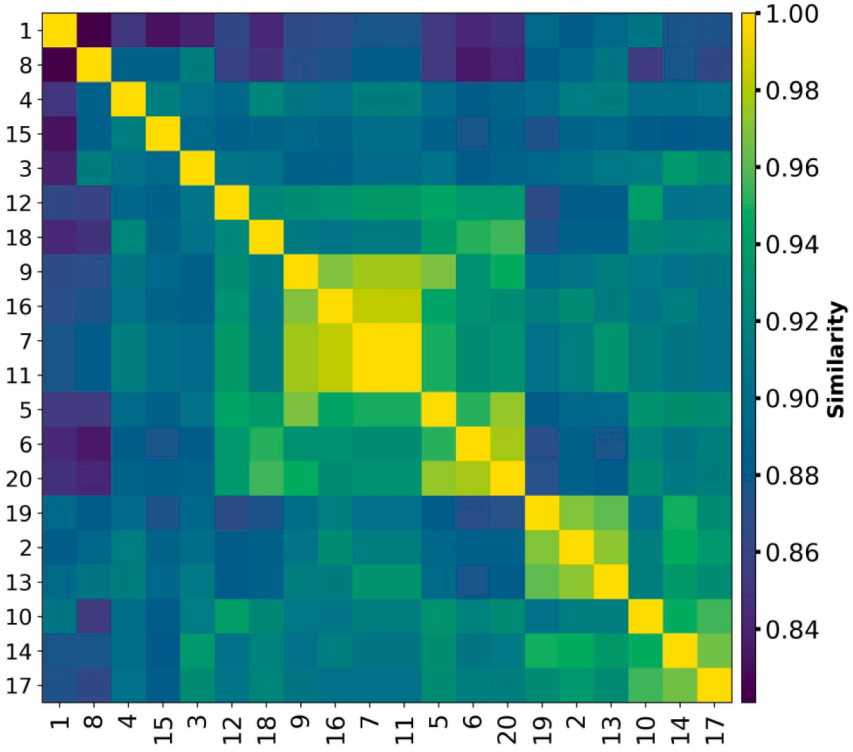


Fig. 11. Heatmap of similarity between OSCs.

CRediT authorship contribution statement

Khadijah Mohammedsaleh Katubi: Writing – original draft, Investigation, Funding acquisition. **Alvi Muhammad Rouf:** Writing – review & editing, Supervision. **Bilal Siddique:** Writing – review & editing, Conceptualization. **Muhammad Faizan Nazar:** Writing – review & editing, Supervision. **Ghulam Jillani Ansari:** Data curation, Visualization. **Z.A. Alrowaili:** Writing – review & editing. **M.S. Al-Buriahi:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

The authors express their gratitude to Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R26), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] M. Sulaman, S. Yang, A. Bukhtiar, C. Fu, T. Song, H. Wang, Y. Wang, H. Bo, Y. Tang, B. Zou, High performance solution-processed infrared photodetector based on PbSe quantum dots doped with low carrier mobility polymer poly(N-vinylcarbazole), RSC Adv. 6 (2016) 44514–44521.

[2] M. Sulaman, Y. Song, S. Yang, Q. Hao, Y. Zhao, M. Li, M.I. Saleem, P. V. Chandraseakar, Y. Jiang, Y. Tang, B. Zou, High-performance solution-processed colloidal quantum dots-based tandem broadband photodetectors with dielectric interlayer, Nanotechnology 30 (2019) 465203.

[3] G. Luo, J. Xie, J. Liu, Y. Luo, M. Li, Z. Li, P. Yang, L. Zhao, K. Wang, R. Maeda, Z. Jiang, Highly stretchable, knittable, wearable fiberform hydrovoltaic generators driven by water transpiration for portable self-power supply and self-powered strain sensor, Small 20 (2024) 2306318.

[4] W. Zhang, Q. Fu, J. Li, B. Lian, Y. Xia, L. Zhou, A.T.S. Wee, P.K.J. Wong, Probing van der Waals magnetic surface and interface via circularly polarized X-rays, Appl. Phys. Rev. 10 (2023).

[5] M. Sulaman, S. Yang, T. Song, H. Wang, Y. Wang, B. He, M. Dong, Y. Tang, Y. Song, B. Zou, High performance solution-processed infrared photodiode based on ternary PbSxSe1−x colloidal quantum dots, RSC Adv. 6 (2016) 87730–87737.

[6] M.I. Abdullah, M.R.S.A. Janjua, M.F. Nazar, A. Mahmood, Quantum chemical designing of efficient TC4-based sensitizers by modification of auxiliary donor and π-spacer, Bull. Chem. Soc. Jpn. 86 (2013) 1272–1281.

[7] D. Wang, X.-X. Wang, M.L. Jin, P. He, S. Zhang, Molecular level manipulation of charge density for solid-liquid TENG system by proton irradiation, Nano Energy 103 (2022) 107819.

[8] M.R.S.A. Janjua, A. Mahmood, F. Ahmad, Solvent effects on nonlinear optical response of certain tetramminеruthenium(II) complexes of modified 1,10-phenanthrolines, Can. J. Chem. 91 (2013) 1303–1309.

[9] C.-H. Li, Z. Huang, J. Lin, T. Hou, Y. Zi, J. Li, Excellent-moisture-resistance fluorinated polyimide composite film and self-powered acoustic sensing, ACS Appl. Mater. Interfaces 15 (2023) 35459–35468.

[10] M.R. Janjua, A. Mahmood, M.F. Nazar, Z. Yang, S. Pan, Electronic absorption spectra and nonlinear optical properties of ruthenium acetylide complexes: a DFT study toward the designing of new high NLO response compounds, Acta Chim. Slov. 61 (2014) 382–390.

[11] M.R.S.A. Janjua, How does bridging core modification alter the photovoltaic characteristics of triphenylamine-based hole transport materials? Theoretical understanding and prediction, Chem. Eur. J. 27 (2021) 4197–4210.

[12] S.-U.-D. Khan, A. Mahmood, U.A. Rana, S. Haider, Utilization of electron-deficient thiadiazole derivatives as π-spacer for the red shifting of absorption maxima of diarylamine-fluorene based dyes, Theor. Chem. Acc. 134 (2014) 1596.

[13] M.R.S.A. Janjua, Quantum chemical design of D–π–A-type donor materials for highly efficient, Photostable, and Vacuum-Processed Organic Solar Cells, Energy Technol. 9 (2021) 2100489.

[14] A. Mahmood, J.-L. Wang, A time and resource efficient machine learning assisted design of non-fullerene small molecule acceptors for P3HT-based organic solar cells and green solvent selection, J. Mater. Chem. A 9 (2021) 15684–15695.

[15] M.R.S.A. Janjua, S. Jamil, A. Mahmood, A. Zafar, M. Haroon, H.N. Bhatti, Solvent-dependent non-linear optical properties of 5,5′-disubstituted-2,2′-bipyridine complexes of ruthenium(<>ii< >): a quantum chemical perspective, Aust. J. Chem. 68 (2015) 1502–1507.

[16] M.R.S.A. Janjua, Theoretical understanding and role of guest π-bridges in triphenylamine-based donor materials for high-performance solar cells, Energy Fuel 35 (2021) 12451–12460.

[17] A. Irfan, A. Mahmood, Computational designing of low energy gap small molecule acceptors for organic solar cells, J. Mex. Chem. Soc. 61 (2017) 309–316.

[18] A. Mahmood, Y. Sandali, J.-L. Wang, Easy and fast prediction of green solvents for small molecule donor-based organic solar cells through machine learning, Phys. Chem. Chem. Phys. 25 (2023) 10417–10426.
[19] A. Mahmood, A. Irfan, J.-L. Wang, machine learning for organic photovoltaic polymers: a minireview, Chin. J. Polym. Sci. 40 (2022) 870–876.
[20] G. Landrum, RDKit: Open-source cheminformatics. http://www.rdkit.org.
[21] D. Rogers, M. Hahn, Extended-connectivity fingerprints, J. Chem. Inf. Model. 50 (2010) 742–754.
[22] I. Muhammad, A. Ali, L. Zhou, W. Zhang, P.K.J. Wong, Vacancy-engineered half-metallicity and magnetic anisotropy in CrSI semiconductor monolayer, J. Alloys Compd. 909 (2022) 164797.
[23] N. Ahmad, H. Zhou, P. Fan, G. Liang, Recent progress in cathode interlayer materials for non-fullerene organic solar cells, EcoMat 4 (2022) e12156.
[24] J. Degen, C. Wegscheid-Gerlach, A. Zaliani, M. Rarey, On the art of compiling and using 'drug-like' chemical fragment spaces, ChemMedChem 3 (2008) 1503–1507.
[25] A. Mahmood, J.-L. Wang, Machine learning for high performance organic solar cells: current scenario and future prospects, Energy Environ. Sci. 14 (2021) 90–105.
[26] M.R.S.A. Janjua, Impact of symmetry breaking on the performance of non-fullerene acceptors (NFAs) for photo and thermally stable organic solar cells (OSCs): A DFT-based interrogation and investigation, J. Photochem. Photobiol. A 444 (2023) 115003.
[27] M.R.S.A. Janjua, All-small-molecule organic solar cells with high fill factor and enhanced open-circuit voltage with 18.25 % PCE: Physical insights from quantum chemical calculations, Spectrochim. Acta - a: Mol. Biomol. Spectrosc. 279 (2022) 121487.
[28] Y. Luo, G. Chen, S. Chen, N. Ahmad, M. Azam, Z. Zheng, Z. Su, M. Cathelinaud, H. Ma, Z. Chen, P. Fan, X. Zhang, G. Liang, Carrier transport enhancement mechanism in highly efficient antimony selenide thin-film solar cell, Adv. Funct. Mater. 33 (2023) 2213941.
[29] M.R.S.A. Janjua, Photovoltaic properties and enhancement in near-infrared light absorption capabilities of acceptor materials for organic solar cell applications: A quantum chemical perspective via DFT, J. Phys. Chem. Solids 171 (2022) 110996.
[30] W. Kuang, H. Wang, X. Li, J. Zhang, Q. Zhou, Y. Zhao, Application of the thermodynamic extremal principle to diffusion-controlled phase transformations in Fe-C-X alloys: Modeling and applications, Acta Mater. 159 (2018) 16–30.
[31] M.R.S.A. Janjua, Quantum chemical design and prediction that complements understanding: How do the transition metals enhance the CO2 sensing ability of inorganic Mg12O12 nanoclusters? J. Phys. Chem. Solids 167 (2022) 110789.
[32] T. Yue, K. Li, X. Li, N. Ahmad, H. Kang, Q. Cheng, Y. Zhang, Y. Yue, Y. Jing, B. Wang, S. Li, J. Chen, G. Huang, Y. Li, Z. Fu, T. Wu, S.U. Zafar, L. Zhu, H. Zhou, Y. Zhang, A binary solution strategy enables high-efficiency quasi-2d perovskite solar cells with excellent thermal stability, ACS Nano 17 (2023) 14632–14643.
[33] A. Mahmood, A. Irfan, J.-L. Wang, Developing efficient small molecule acceptors with sp2-hybridized nitrogen at different positions by density functional theory calculations, molecular dynamics simulations and machine learning, Chem. Eur. J. 28 (2022) e202103712.
[34] X. Chen, Y. Zhao, N. Ahmad, J. Zhao, Z. Zheng, Z. Su, X. Peng, X. Li, X. Zhang, P. Fan, G. Liang, S. Chen, Achieving high open-circuit voltage in efficient kesterite solar cells via lanthanide europium ion induced carrier lifetime enhancement, Nano Energy 124 (2024) 109448.