ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Generation of Chemical Space of Compounds for Prostate Cancer Treatment: Biological Activity Prediction, Clustering, and Visualization of Chemical Space

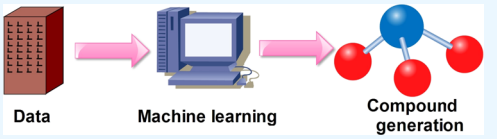
Muhammad Ishfaq,* Mohamed Ibrahim Halawa, Ashfaq Ahmad, Aamir Rasool, Robina Manzoor, Kaleem Ullah, and Yurong Guan*

Cite This: *ACS Omega* 2023, 8, 39408−39419

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Designing molecules for pharmaceutical purposes has been a significant focus for several decades. The pursuit of novel drugs is an arduous and financially demanding undertaking. Nevertheless, the integration of computer-assisted frameworks presents a swift avenue for designing and screening drug-like compounds. Within the context of this research, we introduce a comprehensive approach for the design and screening of compounds tailored to the treatment of prostate cancer. To forecast the biological activity of these compounds, we employed machine learning (ML) models. Additionally, an automated process involving the deconstruction and reconstruction of molecular building blocks leads to the generation of novel compounds. Subsequently, the ML models were utilized to predict the biological activity of the designed compounds, and the t-SNE method was employed to visualize the chemical space covered by the novel compounds. A meticulous selection process identified the most promising compounds, and their potential for synthesis was assessed, offering valuable guidance to experimental chemists in their investigative endeavors. Furthermore, fingerprint and heatmap analysis were conducted to evaluate the chemical similarity among the selected compounds. This multifaceted approach, encompassing predictive modeling, compound generation, visualization, and similarity assessment, underscores our commitment to refining the process of identifying potential candidates for further exploration in prostate cancer treatment.



1. INTRODUCTION

The prostate gland, an accessory organ of the male reproductive system, is found below the bladder and surrounds the urethra. It plays a vital role in the formulation of ejaculate by producing essential fluids, consequently supporting the health and viability of sperm.[1] Prostate glands generally develop tumors in old age, typically after the age of 50.[2] The prostate gland of the adult male can be classified into the following regions: central, transition, and peripheral zones, as well as fibromuscular and periurethral.[3,4] The peripheral zone covers over 70% of the prostate gland in young adult men; therefore, it holds the largest share when it comes to the execution of the normal function of the prostate gland. It has been reported in various studies that the peripheral zone is the region of the prostate gland where neoplasms are most frequently produced in the aged prostate, and around 80% of cases of prostate cancer have been reported due to the accumulation of neoplasms in this region.[3,4] Prostate cancer is a prevalent malignancy in males, which is a common cause of cancer-related death in men.[5] Annually, prostate cancer affects the health and lifestyle of millions of men around the globe, as reported by numerous studies. Anticipating an individual's disease trajectory often relies on analyzing the histopathological, anatomical, and molecular profiles of the tumor along with the patient's health condition. Prostate cancer is a highly important multidisciplinary research field that encompasses various subjects including computational biology, laboratory research, clinical science, and many more.

The combined potential of artificial intelligence and machine learning holds great promise for the significant progress and advancement in the field of predicting properties and designing molecules.[6,7] Subsequently, numerous methods, tools, and models have been devised that can effectively analyze complex and nonlinear data.[8] The journey of drug discovery and development is intricate and multifaceted, entailing a multitude of intricate variables. Machine learning (ML) methods offer a variety of tools to improve decision-making and facilitate discovery throughout the drug discovery process, particularly for well-defined problems with sufficient high-quality data. Some examples of these applications include clinical trial analysis, target validation, and the development of prognostic biomarkers, among others. Machine learning techniques, such as QSAR analysis, hit discovery, and de novo drug design, enable the pharmaceutical industry to make informed decisions

Received: July 13, 2023
Accepted: October 3, 2023
Published: October 16, 2023

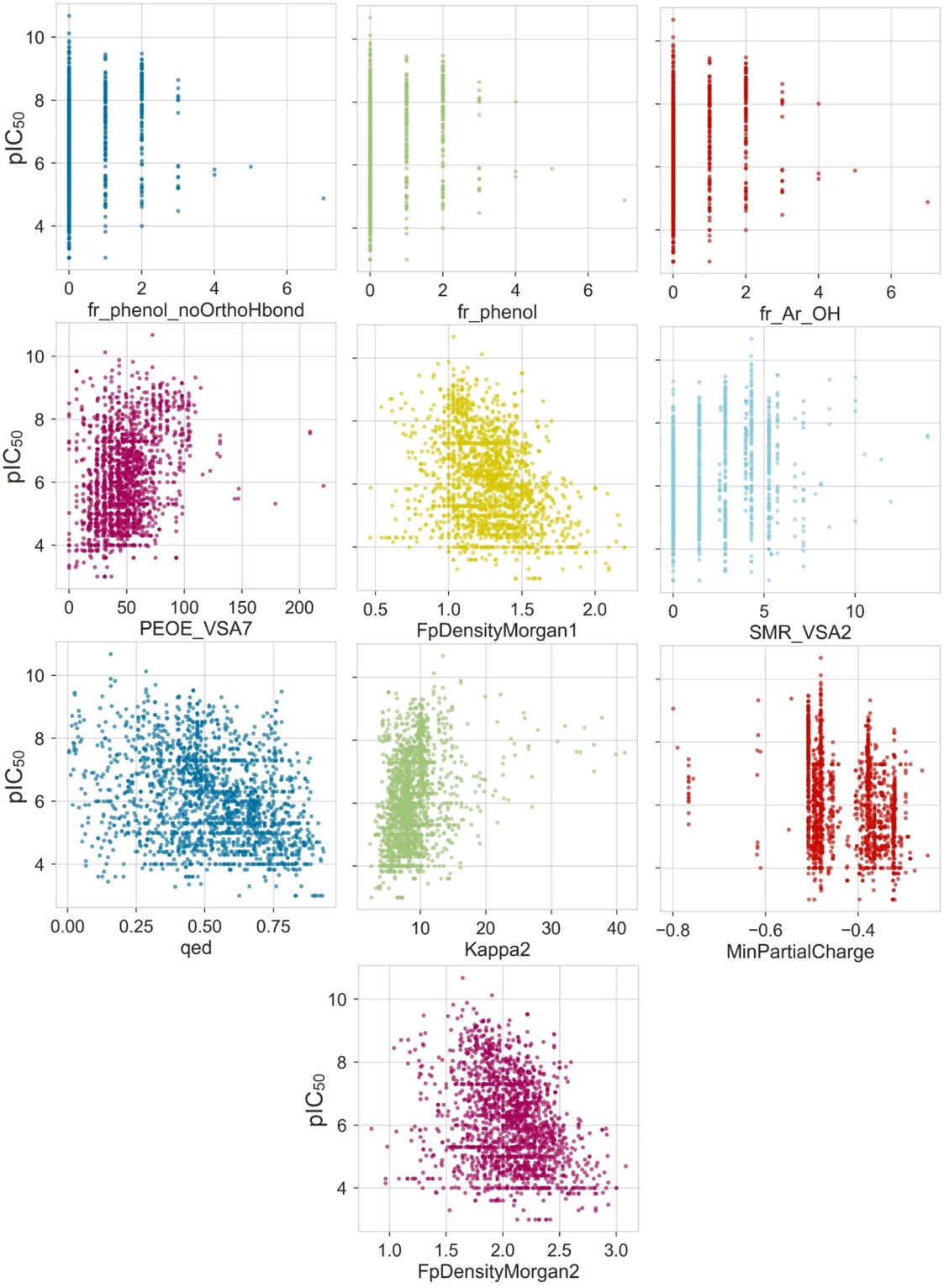


Figure 1. Scatter plots between pIC_{50} and the top descriptors.

and achieve better outcomes.[9] As a result, these methodologies are currently driving the achievements of experts within the pharmaceutical industry.[10,11] Deep structured learning has risen as an innovative machine learning approach with substantial ramifications across scientific domains, especially in cases where the intricate nature of biological systems eludes comprehensive modeling through physical-based methods, necessitating a more sophisticated approach.

The pharmaceutical sector has embraced the realm of ML initiatives, embarking on the integration of this technology within the drug development processes. The evolution of ML methodologies, coupled with the burgeoning repository of pharmacological data, has elevated the significance of artificial intelligence (AI) in shaping the landscape of drug design. A pivotal facet of AI lies in its capability to translate intricate medical data into reproducible approaches, thereby diminish-

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

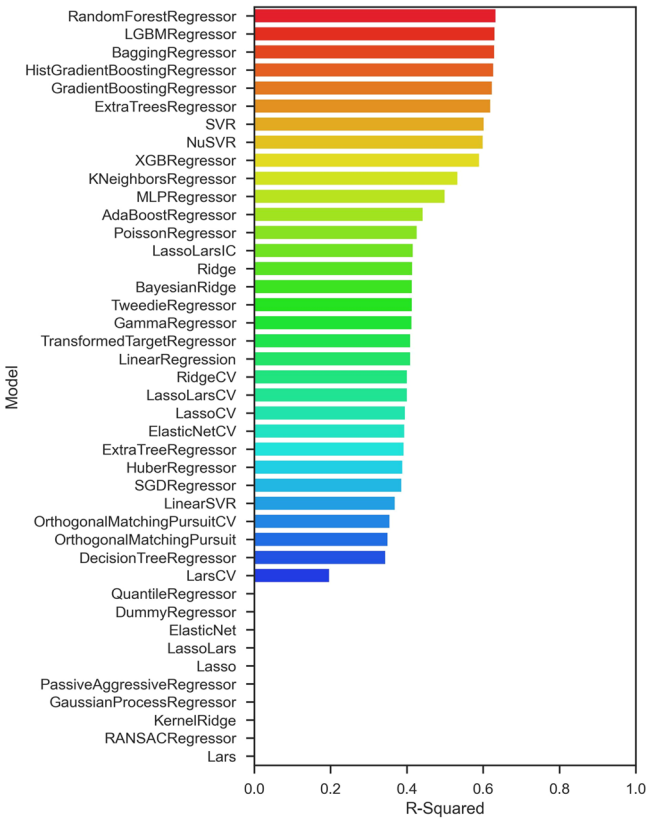


Figure 2. Performance comparison of various machine learning models for the test set.

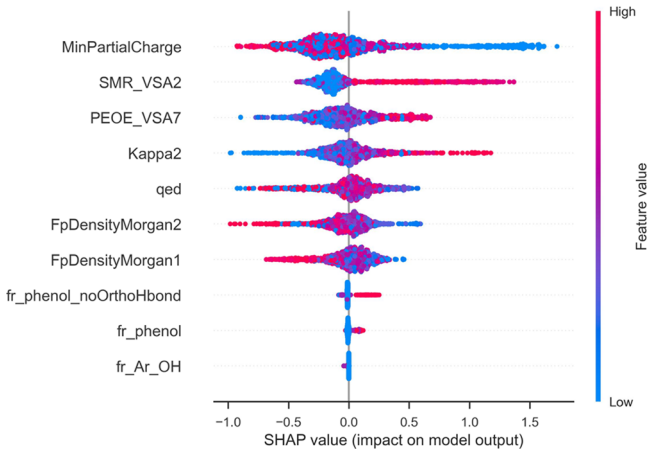


Figure 4. Positive and negative impacts of selected descriptors on the output of random forest regressor (fr_phenol_noOrthoHbond, fr_phenol, and fr_Ar_OH are low contributing features).

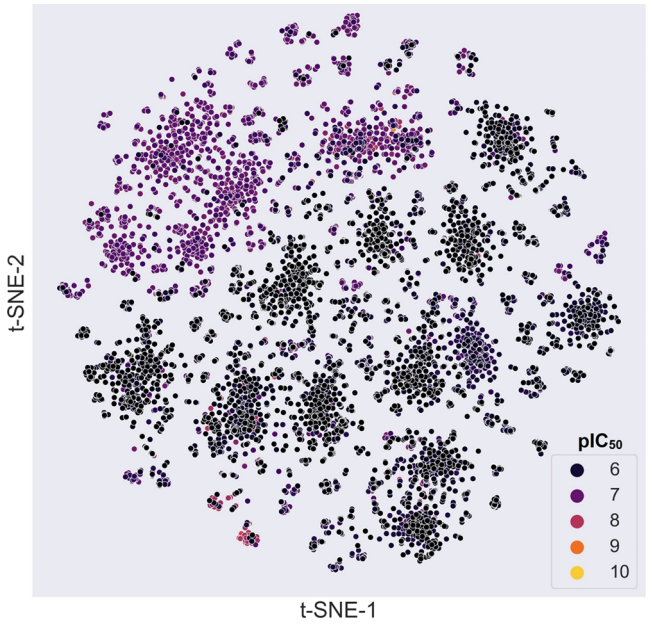


Figure 5. t-SNE plot.

ing the need for speculative enhancements. Within this context, a pivotal contribution emerges through the establishment of a stability model for drug sensitivity. This model is crafted utilizing ML techniques orchestrated by dissecting preclinical data. Subsequently, the efficacy of this model is tested by using clinical samples from patients, ushering in a phase of validation to ascertain its accuracy. This, in turn, offers invaluable insights into disease indications and facilitates a streamlined selection process, expediting the course of clinical drug development. The burgeoning potential of ML

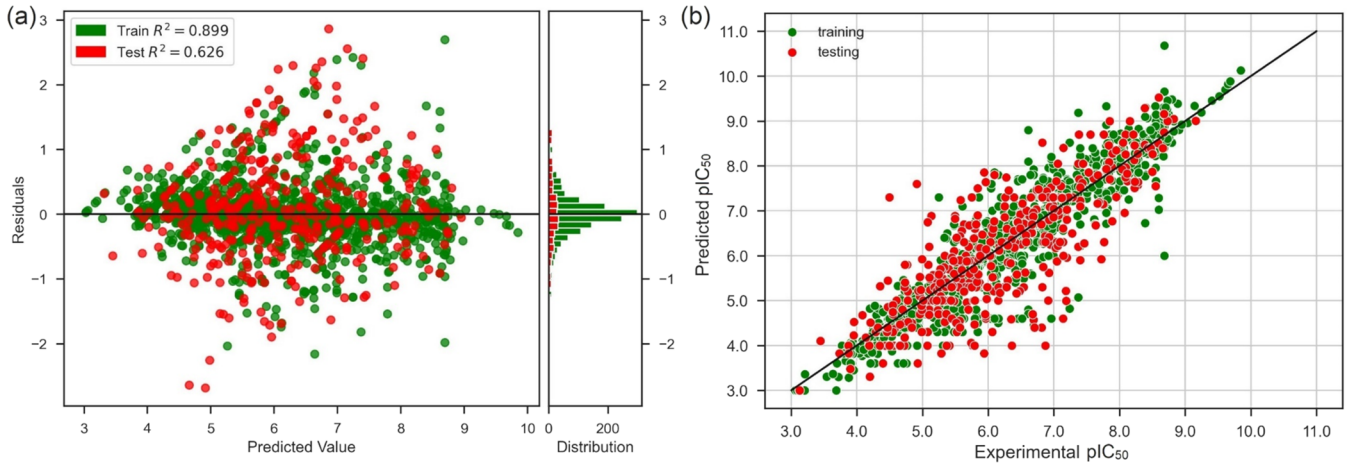


Figure 3. (a) Residuals for random forest regressor. (b) Scatter plot between experimental and predicted pIC50 values using random forest regressor.

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

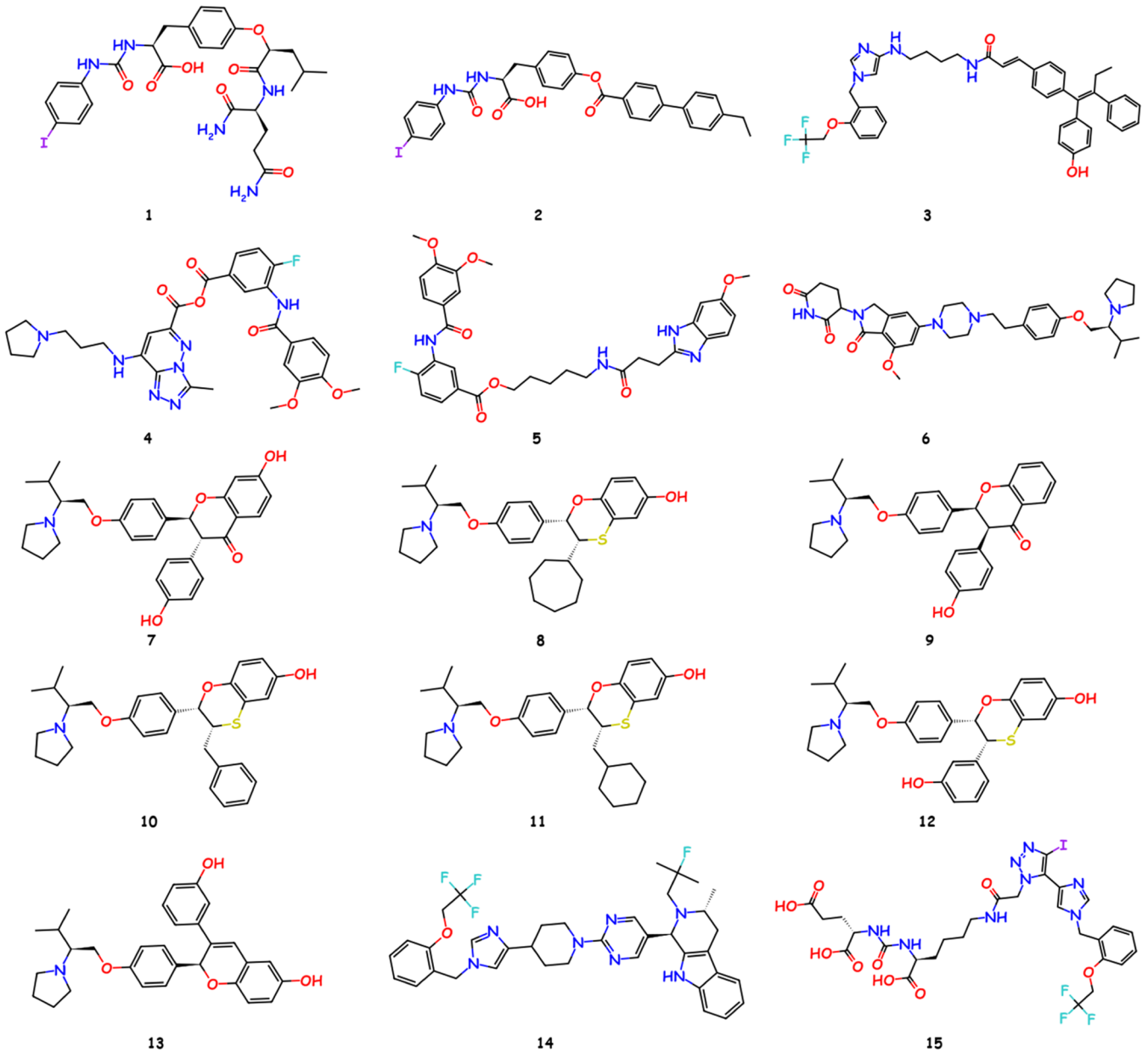


Figure 6. Chemical structures of selected compounds 1−15.

methodologies is poised to supplant traditional practices within this domain, ushering in a transformative era of advancement.

In this research, our focus was on designing and evaluating compounds for prostate cancer treatment. Machine learning models play a vital role in prognosticating their biological activity. The employment of the t-SNE technique unveiled the intricate chemical landscape covered by these newly designed compounds. From this array, standout candidates were meticulously selected, considering factors such as synthetic feasibility. The tapestry of chemical likeness was further explored through cluster analysis and heatmap visualization, enhancing our understanding of compound relationships.

2. MATERIALS AND METHODS

2.1. Machine Learning Analysis. The data for machine learning is collected from the CHEMBL database.[12] Search is done using the word "prostate cancer". Only the data of a single protein are sorted out, and the protein complex is not considered. To obtain more data, all protein targets are considered. The data of 2000 compounds are used for machine learning analysis. Half-maximal inhibitory concentration (IC_{50}) values are converted into pIC_{50}, that is, the negative log of IC_{50}. The distribution of the pIC_{50} values is given in Figure S1. The structure of 10 compounds with lowest pIC_{50} values and 10 compounds with highest pIC_{50} values is given in Figure S2. The molecular descriptors are calculated using RDKit.[13] It generates about 200 descriptors. Pandas module was used for importing the data library with determined optimum descriptors and target property in the comma-isolated (.CSV) format. Seaborn, Matplotlib, and Scikit-learn are used for machine learning and data visualization. The interpretation of ML models is done using SHapley Additive exPlanations (SHAP).[14] It helps to understand the impact of different features (descriptors) on the output of ML models.

2.2. Compound Design and Similarity Analysis. The breaking retrosynthetically interesting chemical substructures

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

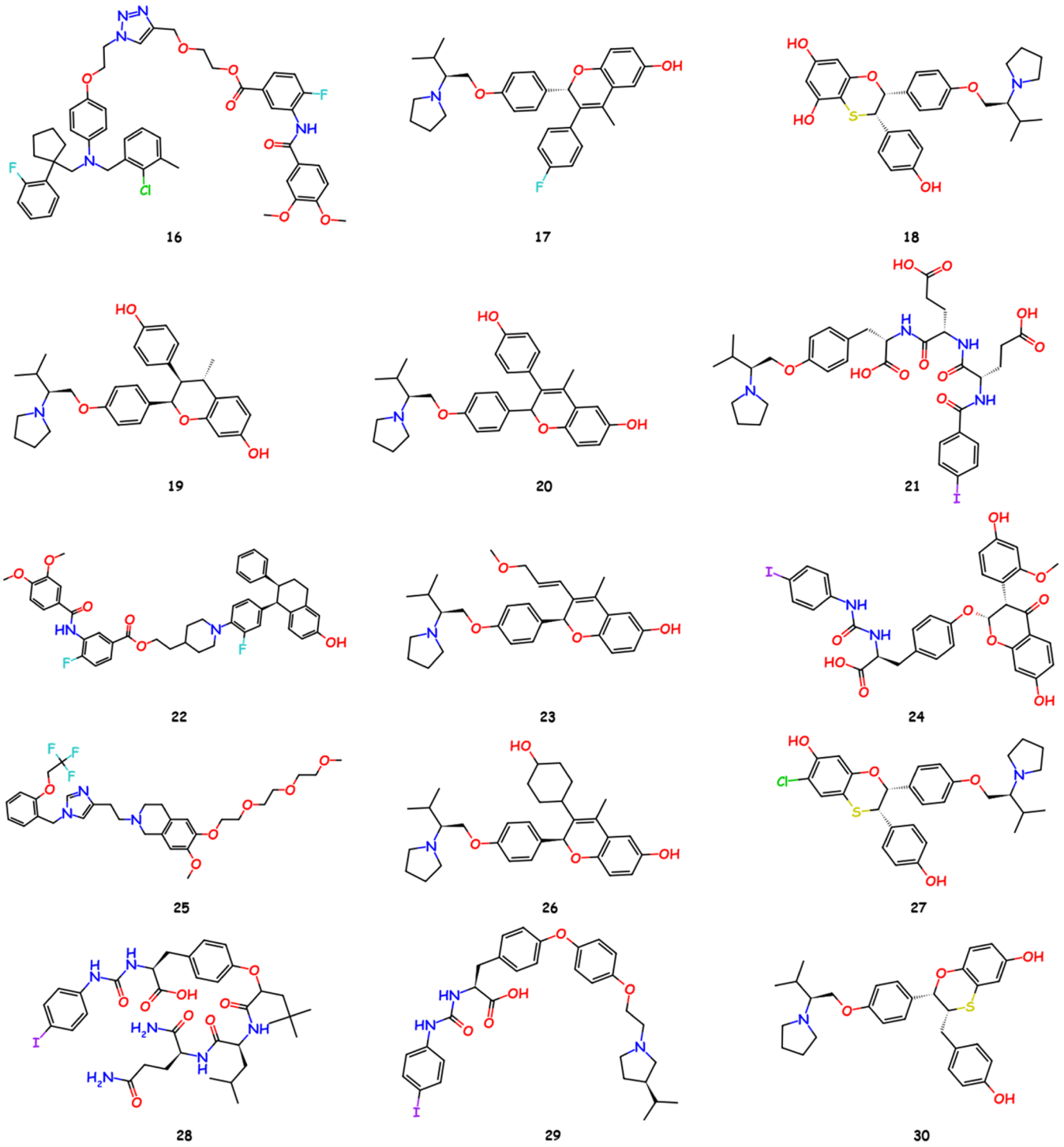


Figure 7. Chemical structures of selected compounds 16−30.

(BRICS) method[15] is used to design new compounds. BRICS is implanted on RDKit.[13] A set of compounds as input is required. The 500 compounds in the data set with the highest pIC50 values are selected as input. BRICS breaks the compounds into fragments and then joins them to design new compounds. We have generated 20,000 compounds. The pIC50 values of the generated compounds are predicted using the already trained ML model. Then, the generated chemical space is visualized using a t-distributed stochastic neighbor embedding (t-SNE) plot. Designed compounds are shortlisted on the basis of predicted values. Chemical similarity and clustering are performed on selected compounds. For this purpose, chemical fingerprints are used. Synthetic accessibility is also calculated.

3. RESULTS AND DISCUSSION

3.1. Machine Learning Analysis. Molecular descriptors can be described as either experimental or computationally derived values that are associated with a specific molecule.[16,17] Alternatively, a molecular descriptor is the outcome of a logical and mathematical transformation that converts chemical information into a numerical representation or a standardized

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

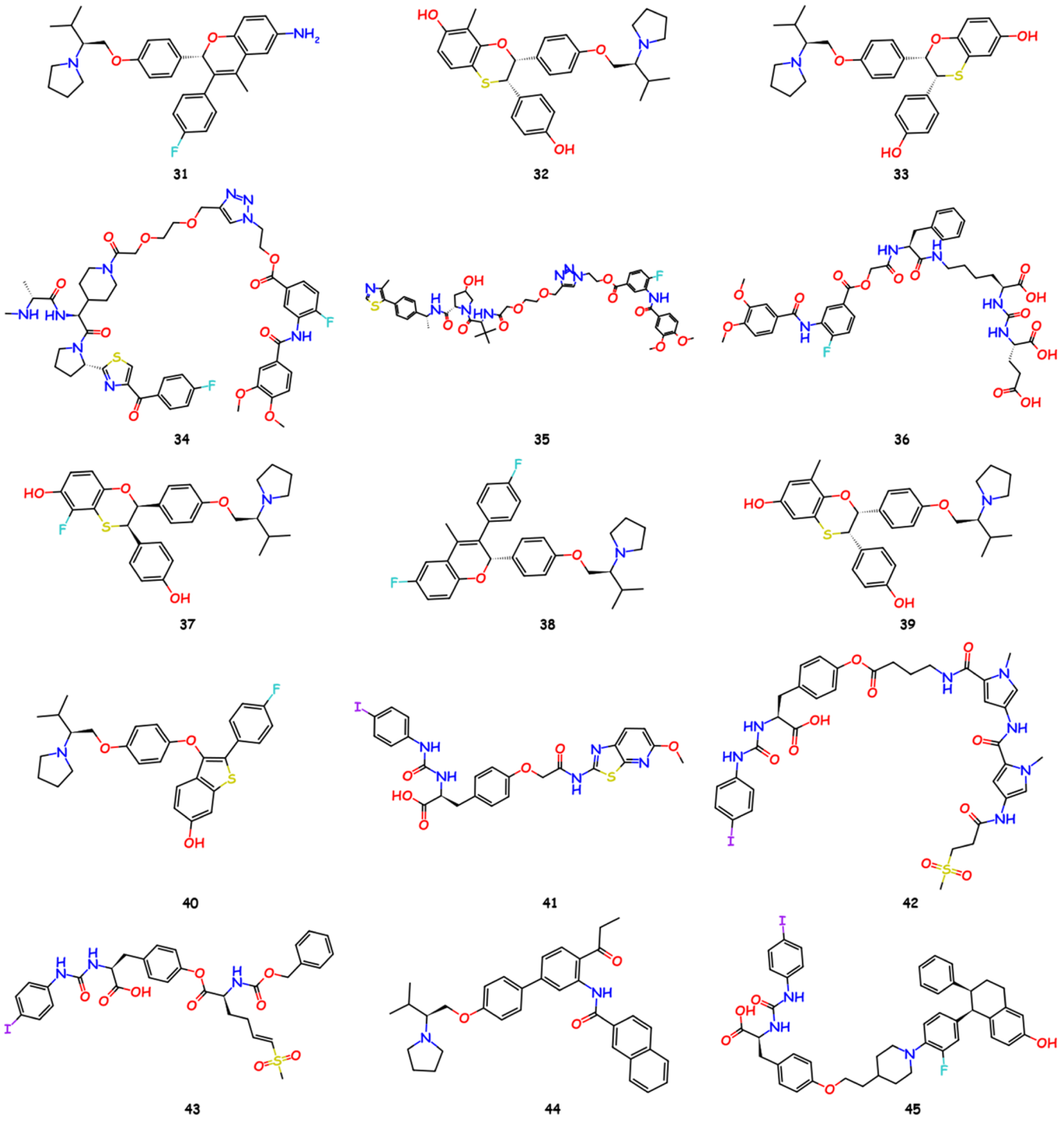


Figure 8. Chemical structures of selected compounds 31−45.

experimental result.[18,19] These descriptors enable qualitative and quantitative analyses of chemical data. Some examples of molecular descriptors include structural, topological, electronic, and physicochemical descriptors.

Molecular descriptors encompass various quantitative representations of molecules. These descriptors prove valuable in conducting similarity searches within molecular libraries, enabling the identification of molecules with similar physical or chemical properties based on shared descriptor values.[20,21] Furthermore, molecular descriptors play a crucial role in prediction models. They establish a correlation between the structure−property relationship and aid in predicting the properties of molecules by considering their descriptor values. The literature contains a wide range of molecular descriptors, spanning from simple bulk properties to complex three-dimensional formulations and extensive molecular fingerprints with thousands of bit positions. Selecting the most suitable descriptors for specific applications based on knowledge is an important task in chemoinformatics research.[22] To ensure rational selection rather than relying on guesses or chemical intuition, a thorough evaluation of the descriptor performance is necessary. Figure 1 shows the correlation plots between pIC_{50} and the top descriptors. From the figure, it is clear that individually most of the descriptors are not showing a strong

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

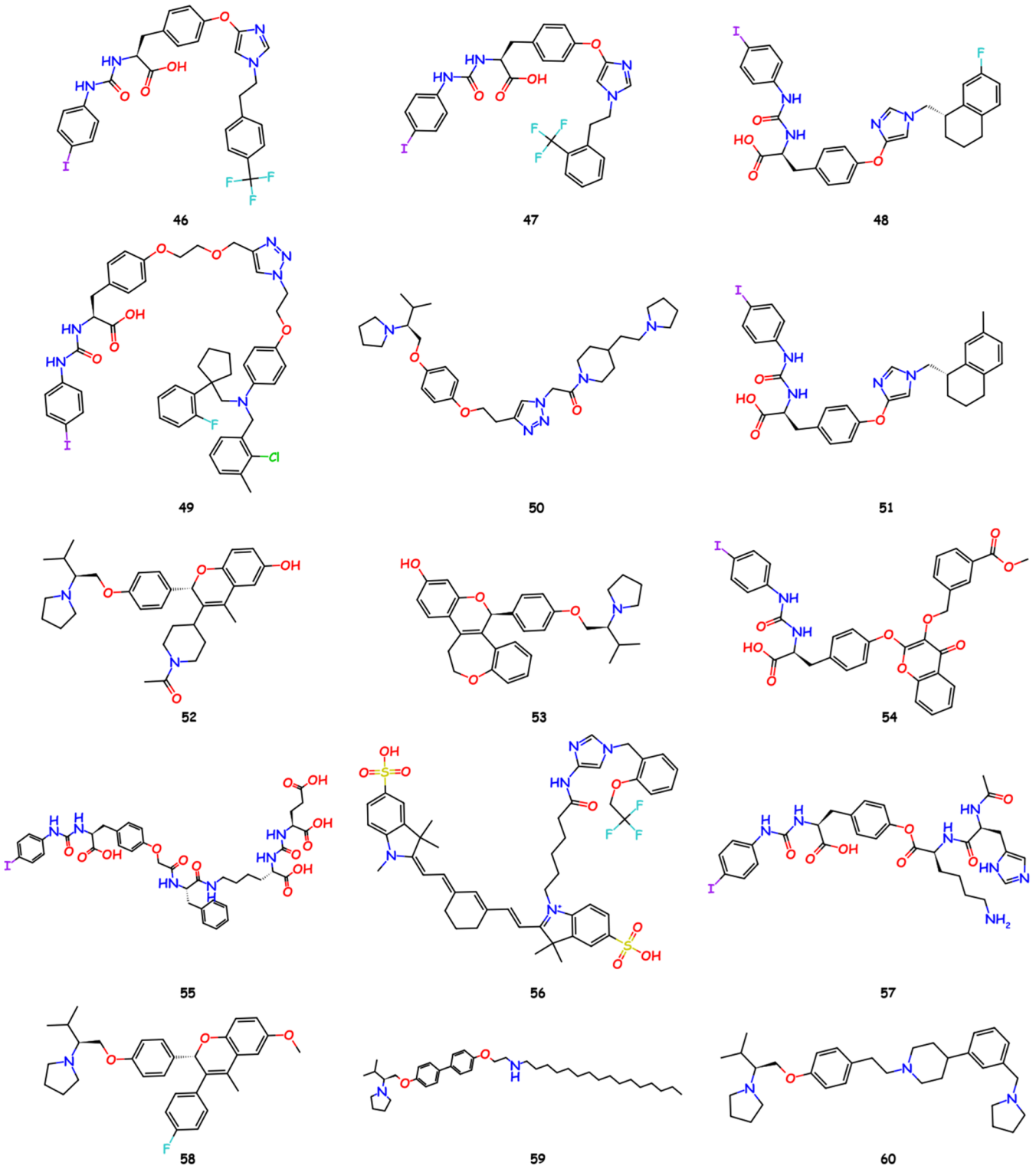


Figure 9. Chemical structures of selected compounds 46−60.

correlation with target property. Jointly, all of the descriptors are useful in model training.

Machine learning is a field that explores the construction of computer systems capable of improving through experience while uncovering the fundamental laws governing learning systems.[23,24] It has rapidly evolved from a curiosity to a practical technology in the past two decades, finding extensive use in commercial applications and becoming the preferred approach in AI for tasks such as computer vision and natural language processing. Machine learning's impact extends beyond AI, influencing various domains, including data-intensive fields like consumer services and fault diagnosis.[25,26] Moreover, empirical sciences have benefited from machine learning's ability to analyze high-throughput experimental data, revolutionizing disciplines such as biology, cosmology, and social science.

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

Table 1. pIC_{50} and Synthetic Accessibility Score of Selected Compounds 1−30

name	pIC_{50}	synthetic accessibility score
1	10.637	3.522
2	10.565	2.673
3	10.507	3.148
4	10.503	2.882
5	10.503	2.546
6	10.482	3.741
7	10.444	3.614
8	10.444	3.805
9	10.444	3.481
10	10.444	3.681
11	10.444	3.856
12	10.444	3.72
13	10.444	3.525
14	10.421	4.271
15	10.421	4.058
16	10.399	3.552
17	10.395	3.433
18	10.395	3.874
19	10.395	3.801
20	10.395	3.478
21	10.389	3.926
22	10.374	3.707
23	10.371	3.758
24	10.371	3.79
25	10.369	3.003
26	10.366	3.665
27	10.366	3.806
28	10.361	4.023
29	10.36	3.178
30	10.356	3.78

Table 2. pIC_{50} and Synthetic Accessibility Score of Selected Compounds 31−60

name	pIC_{50}	synthetic accessibility score
31	10.356	3.437
32	10.356	3.798
33	10.334	3.682
34	10.328	4.774
35	10.328	4.797
36	10.319	3.87
37	10.299	3.865
38	10.299	3.387
39	10.299	3.826
40	10.282	3.055
41	10.28	2.972
42	10.274	3.658
43	10.265	3.539
44	10.256	2.935
45	10.254	3.952
46	10.25	3.15
47	10.25	3.214
48	10.25	3.59
49	10.243	4.032
50	10.232	3.274
51	10.229	3.577
52	10.229	3.651
53	10.215	3.594
54	10.21	3.237
55	10.21	4.087
56	10.208	4.28
57	10.182	3.715
58	10.176	3.381
59	10.172	3.019
60	10.172	2.928

Globally, gastric cancer has one of the highest mortality rates among cancers with a current survival rate of only 30% even when using combination therapies. However, recent evidence suggests that miRNAs (microRNAs) may have a potential role in diagnosing and assessing the prognosis of various cancers, including gastric cancer.[27] In the field of cancer research, machine learning (ML) has become increasingly prominent as a tool for identifying clinically relevant biomarkers with practical applications. Biofilm production in bacteria contributes to the severity of infections and poses challenges for antimicrobial treatment. Bacteriophage depolymerases, the enzymes employed by viruses, offer a potential solution to degrade the biofilm matrix. The machine learning-based method accurately identifies phage depolymerases using a limited set of validated enzymes, which highlights the potential machine learning for protein functional annotation and the discovery of novel therapeutic agents.[28]

There are so many machine learning models. Their performance depends on the used data set. The machine learning model is not suitable for every type of data. A specific machine learning model is of high significance because it determines the output (results).[29,30] Therefore, Lazy Predict (a Python-based code) is used to test around 40 machine learning models.[31] For this analysis, the data set is divided into a 70:30% train:test ratio. The bar graph based on the R-squared values for the test set is given in Figure 2. Only a few models show a higher performance. The random forest model is the best model among all tested models. This model is selected for further analysis. For the best model, various ratios of the training and test sets are tried. The 70:30 ratio is the best. The hyperparameters of random forest models are also optimized. However, no significant difference is observed. The residual plot for random forest regressor is given in Figure 3(a). R-squared values for training and test sets are 0.899 and 0.626, respectively. The model is responsibly accurate. Values of residuals are not high. The scatter plot between the predicted and true values for the training and test sets using the random forest model is given in Figure 3(b). Most points are near the standard line.

To further get insights about the ML working model and identify the possible contributions of features on the output of the model, SHapley Additive exPlanations (SHAP) was applied.[14] The SHAP value ranks the samples by allocating them a specific number based on the selection of optimum sample features in comparison with the desired features. Positive and negative impacts of selected descriptors on the output of random forest regressor are given in Figure 4. For MinPartialCharge, most low values have a high positive impact on the model. High values of these descriptors have a low negative impact on model performance. SMR_VSA2 has shown opposite behavior. In the case of PEOE_VSA7 and Kappa2, higher values of descriptors have shown positive impact and negative values have shown negative impact. In the case of the QED descriptor, behavior is mixed. Other descriptors have shown mixed behavior or low impact.

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

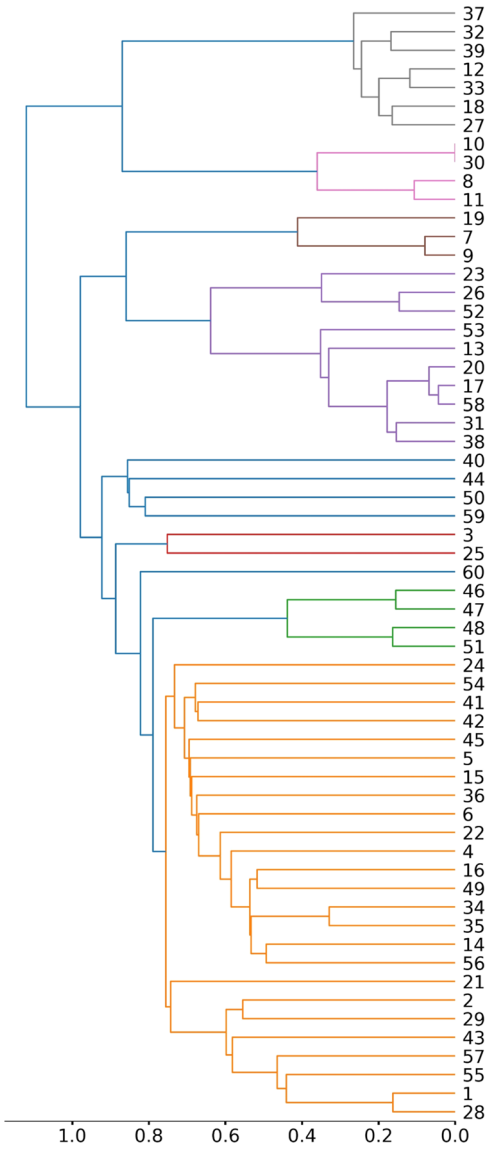


Figure 10. Clustering of compounds on the basis of similarity.

3.2. Designing of New Compounds. De novo molecular design is a challenging task aimed at creating new chemical compounds with specific properties and activities.[32] The vast search space of feasible molecules, estimated to be around 10^{33}–10^{80}, poses a significant challenge as only a small fraction exhibits the desired traits.[33] Traditionally, de novo molecular design heavily relies on trial and error, with human expertise and intuition playing a major role. However, the high costs associated with developing new molecules have prompted the use of computational tools to assist in the process, leading to practical applications and widespread adoption in the field. These computational tools have proven to be valuable in mitigating the challenges of de novo molecular design.

The availability of multiple types of modification of organic molecules allows us to design a large number of new molecules.[34,35] The chances to incorporate more heteroatoms also allow us to design more organic semiconductors.[36,37] New building units are also generated using the BRICS method. The 500 compounds in the data set with the highest pIC_{50} are used as input. BRICS analysis is done using RDKit. The BRICS algorithm automatically breaks the compounds into fragments on the basis of predefined rules and then joins these fragments to design new compounds. We have generated 20,000 new compounds. The pIC_{50} values of the generated compounds are predicted using random forest regressor. The distribution of the predicted pIC_{50} values is given in Figure S3. The predicted values are found in a wide range. Majority of compounds have shown values near to 6. Only limited compounds have shown values near to 10.

The generated chemical space of the compounds is visualized using t-SNE. t-SNE is a powerful visualization tool to group the probabilities based on their similarities. Furthermore, it also reduces the noise in high-dimensional data with a huge number of features. The distance between compounds indicates the similarity between compounds. The closer the compounds, the more similar they are. A large number of small patches indicates that these compounds are structurally diverse in nature (Figure 5). The higher and lower values are almost equally distributed.

The generated compounds are screened on the basis of the predicted pIC_{50} values. We selected 60 compounds. Their structures are given in Figures 6–9. The selected compounds are structurally dispersed in nature. Our approach is valuable in finding unique compounds.

3.3. Synthetic Accessibility. The synthesis of compounds requires multiple steps depending on the availability of starting materials.[38,39] The ease of the synthesis is controlled by various factors. The synthetic accessibility score (SAS) is a measure of the ease with which a molecule can be synthesized. It considers all of the possible factors, including availability and cost of starting materials, number of synthetic steps involved, and the possibility of side reactions taking place during synthesis. We have calculated the synthetic accessibility score using RDkit. Results are given in Tables 1 and 2. Synthetic accessibility score values fall between 1 (easy to synthesize) and 10 (difficult to synthesize). Six is considered a threshold to distinguish between easy to synthesize and difficult to synthesize.[40] All of the selected compounds have synthetic accessibility score values lower than 5. It indicates that these compounds are easy to synthesize.

3.4. Similarity Analysis. Clustering of compounds is a tool to categorize compounds together based on their similarity. This technique is helpful in the design and synthesis of new compounds. It also helps experimental scientists to search and identify the compounds with like properties.[39,41] The clustering of compounds on the basis of structural similarity is shown in Figure 10. From the figure, it is clear that compounds are much different. Only a small group of compounds is similar. To further verify the similarity between the compounds, we visualize the similarity between the compounds with the help of a heatmap (Figure 11). It provides the pairwise similarity between compounds.[42] Only a few compounds have similarity higher than 0.9. Majority of compounds have similarity near to 4.

Machine learning models are trained for the prediction of biological activity. A large library of new compounds is designed using an automatic method. In the majority of reported studies, only property prediction is done, and designing and screening of new compounds make our study unique.

4. CONCLUSIONS

Efficient and cost-effective approaches have become crucial in drug design, as they swiftly pinpoint the most promising

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

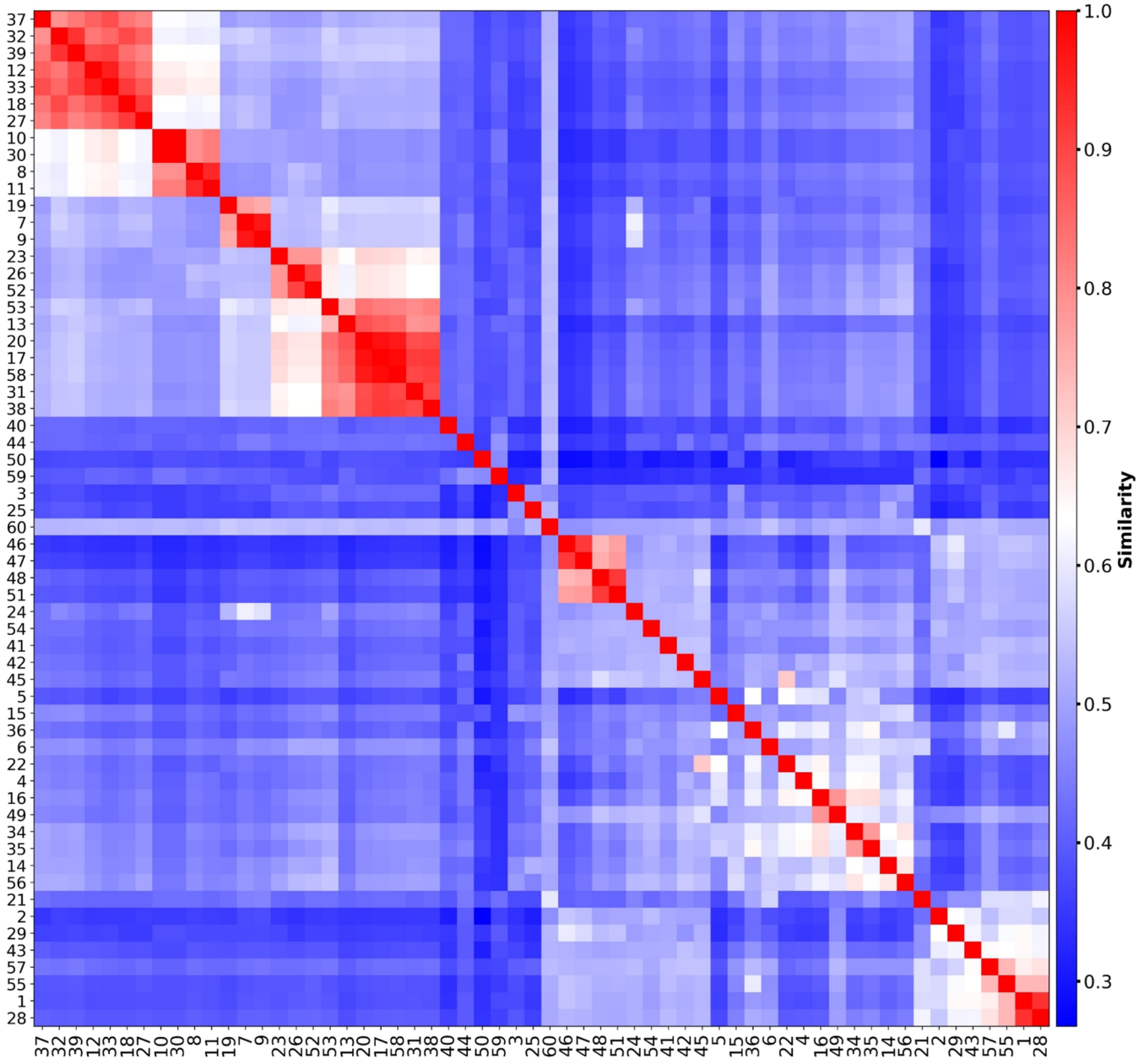


Figure 11. Heatmap of similarity between compounds.

compounds. This research delves into the realm of machine learning, harnessing molecular descriptors to forecast the biological properties of compounds. By leveraging the breaking retrosynthetically interesting chemical substructures (BRICS) technique, a pool of 20,000 novel compounds was synthesized. These newly generated compounds were subjected to established machine learning models that had been previously trained. These models were employed to prognosticate the biological activity of the synthesized compounds. Notably, the application of the t-SNE method unveiled the remarkable diversity inherent in these new compounds, emphasizing their wide-ranging nature. The calculation of synthetic accessibility is applied to both the standard and selected compounds. It is observed that a significant portion of these compounds exhibits a high level of synthetic feasibility. Utilizing a heatmap based on chemical similarity, it becomes evident that a considerable proportion of the chosen compounds displays remarkable dispersion across the chemical space. This innovative framework that we propose holds the potential to substantially assist in the identification of optimal compounds for advancing prostate cancer treatment strategies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c05056.

Figure S1: distribution of experimental pIC_{50} in the collected data; Figure S2: chemical structures of 10 compounds with lowest pIC_{50} and 10 compounds with highest pIC_{50}; and Figure S3: distribution of predicted pIC_{50} of the generated compounds (PDF)

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419

AUTHOR INFORMATION

Corresponding Authors

Muhammad Ishfaq − *College of Computer Science, Huanggang Normal University, Huanggang 438000, China;* orcid.org/0000-0003-1376-8986; Email: muhammad@hgnu.edu.cn

Yurong Guan − *College of Computer Science, Huanggang Normal University, Huanggang 438000, China;* Phone: 008618580344982; Email: jsjgyr@hgnu.edu.cn

Authors

Mohamed Ibrahim Halawa − *Department of Pharmaceutical Analytical Chemistry, Faculty of Pharmacy, Mansoura University, Mansoura, 35516 Mansoura, Egypt; Guangdong Laboratory of Artificial Intelligence & Digital Economy (SZ), Shenzhen University, Shenzhen 518060, P. R. China;* orcid.org/0000-0001-6358-2691

Ashfaq Ahmad − *Chemistry Department, College of Science, King Saud University, Riyadh 11451, Kingdom of Saudi Arabia*

Aamir Rasool − *Institute of Biochemistry, University of Balochistan, Quetta 87300, Pakistan*

Robina Manzoor − *Department of Biotechnology and Bioinformatics, Lasbella University of Agriculture, Water and Marine Sciences, Uthal 90150, Pakistan*

Kaleem Ullah − *Department of Microbiology, University of Balochistan, Quetta 87300, Pakistan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c05056

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This project was supported by the Huanggang Normal University Project (2042022005, 30120230102). The authors sincerely appreciate funding from the Researchers Supporting Project number (RSPD2023R666), King Saud University, Riyadh, Saudi Arabia.

REFERENCES

(1) Verze, P.; Cai, T.; Lorenzetti, S. The Role of the Prostate in Male Fertility, Health and Disease. *Nat. Rev. Urol.* **2016**, *13*, 379−386.

(2) Attard, G.; Parker, C.; Eeles, R. A.; Schröder, F.; Tomlins, S. A.; Tannock, I.; Drake, C. G.; de Bono, J. S. Prostate Cancer. *Lancet* **2016**, *387*, 70−82.

(3) McNeal, J. E. The Zonal Anatomy of the Prostate. *Prostate* **1981**, *2*, 35−49.

(4) Timms, B. G. Prostate Development: A Historical Perspective. *Differentiation* **2008**, *76*, 565−577.

(5) Sung, H.; Ferlay, J.; Siegel, R. L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: Globocan Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Ca: Cancer J. Clin.* **2021**, *71*, 209−249.

(6) Janjua, M. R. S. A.; Irfan, A.; Hussien, M.; Ali, M.; Saqib, M.; Sulaman, M. Machine-Learning Analysis of Small-Molecule Donors for Fullerene Based Organic Solar Cells. *Energy Technol.* **2022**, *10*, No. 2200019.

(7) Mahmood, A.; Wang, J.-L. 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* **2021**, *9*, 15684−15695.

(8) Rizvi, A. S.; Murtaza, G.; Irfan, M.; Xiao, Y.; Qu, F. Determination of Kynurenine Enantiomers by Alpha-Cyclodextrin, Cationic-Beta-Cyclodextrin and Their Synergy Complemented with Stacking Enrichment in Capillary Electrophoresis. *J. Chromatogr. A* **2020**, *1622*, No. 461128.

(9) Hussain, S.; Ali, S.; Shahzadi, S.; Sharma, S. K.; Qanungo, K.; Altaf, M.; Evans, H. S. Synthesis, Characterization, and Semi-Empirical Study of Organotin(Iv) Complexes with 4-(Hydroxymethyl)Piperidine-1-Carbodithioic Acid: X-Ray Structure of Chlorodimethyl-(4-Hydroxymethyl Piperidine-1-Carbodithioato-S,S′)Tin(Iv). *Phosphorus, Sulfur Silicon Relat. Elem.* **2011**, *186*, 542−551.

(10) Ahmad, F.; Mahmood, A.; Muhmood, T. Machine Learning-Integrated Omics for the Risk and Safety Assessment of Nanomaterials. *Biomater. Sci.* **2021**, *9*, 1598−1608.

(11) Hussain, S.; Ali, S.; Shahzadi, S.; Sharma, S. K.; Qanungo, K.; Bukhari, I. H. Homobimetallic Complexes Containing Sn(Iv) with Acetylene Dicarboxylic Acid: Their Syntheses and Structural Interpretation by Spectroscopic, Semi-Empirical, and Dft Techniques. *J. Coord. Chem.* **2012**, *65*, 278−285.

(12) Mendez, D.; Gaulton, A.; Bento, A. P.; et al. Chembl: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.

(13) Landrum, G. *Rdkit: Open-Source Cheminformatics.* Http://Www.Rdkit.Org.

(14) Lundberg, S. M.; Lee, S.-I. In *A Unified Approach to Interpreting Model Predictions*, Proceedings of the 31st International Conference on Neural Information Processing Systems; Curran Associates Inc.: New York, 2017; pp 4768−4777.

(15) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem.* **2008**, *3*, 1503−1507.

(16) Hussain, S.; Ali, S.; Shahzadi, S.; Tahir, M. N.; Shahid, M. Synthesis, Characterization, Single Crystal Xrd and Biological Screenings of Organotin(Iv) Derivatives with 4-(2-Hydroxyethyl)-Piperazine-1-Carbodithioic Acid. *J. Coord. Chem.* **2016**, *69*, 687−703.

(17) Janjua, M. R. S. A.; Jamil, S.; Mahmood, A.; Zafar, A.; Haroon, M.; Bhatti, H. N. Solvent-Dependent Non-Linear Optical Properties of 5,5′-Disubstituted-2,2′-Bipyridine Complexes of Ruthenium(II): A Quantum Chemical Perspective. *Aust. J. Chem.* **2015**, *68*, 1502−1507.

(18) Irfan, A.; Mahmood, A. Computational Designing of Low Energy Gap Small Molecule Acceptors for Organic Solar Cells. *J. Mex. Chem. Soc.* **2017**, *61*, 309−316.

(19) Hussain, S.; Ali, S.; Shahzadi, S.; Tahir, M. N.; Shahid, M.; Munawar, K. S.; Abbas, S. M. Synthesis, Spectroscopy, Single Crystal Xrd and Biological Studies of Multinuclear Organotin Dicarboxylates. *Polyhedron* **2016**, *117*, 64−72.

(20) Irfan, A.; Mahmood, A. Designing of Efficient Acceptors for Organic Solar Cells: Molecular Modelling at Dft Level. *J. Cluster Sci.* **2018**, *29*, 359−365.

(21) Tahir, M. H.; Mubashir, T.; Shah, T.-U.-H.; Mahmood, A. Impact of Electron-Withdrawing and Electron-Donating Substituents on the Electrochemical and Charge Transport Properties of Indacenodithiophene-Based Small Molecule Acceptors for Organic Solar Cells. *J. Phys. Org. Chem.* **2019**, *32*, No. e3909.

(22) Hussain, S.; Ali, S.; Shahzadi, S.; Tahir, M. N.; Ramzan, S.; Shahid, M. Synthesis, Spectroscopic Characterization, X-Ray Crystal Structure and Biological Activities of Homo- and Heterobimetallic Complexes with Potassium-1-Dithiocarboxylatopiperidine-4-Carboxylate. *Polyhedron* **2016**, *119*, 483−493.

(23) Mahmood, A.; Sandali, Y.; Wang, J.-L. Easy and Fast Prediction of Green Solvents for Small Molecule Donor-Based Organic Solar Cells through Machine Learning. *Phys. Chem. Chem. Phys.* **2023**, *25*, 10417−10426.

(24) Khan, S. U.-D.; Mahmood, A.; Rana, U. A.; Haider, S. Utilization of Electron-Deficient Thiadiazole Derivatives as Π-Spacer for the Red Shifting of Absorption Maxima of Diarylamine-Fluorene Based Dyes. *Theor. Chem. Acc.* **2015**, *134*, 1596.

(25) Mubashir, T.; Hussain Tahir, M.; Altaf, Y.; Ahmad, F.; Arshad, M.; Hakamy, A.; Sulaman, M. Statistical Analysis and Visualization of Data of Non-Fullerene Small Molecule Acceptors from Harvard

Organic Photovoltaic Database. Structural Similarity Analysis with Famous Non-Fullerene Small Molecule Acceptors to Search New Building Blocks. *J. Photochem. Photobiol., A* **2023**, *437*, No. 114501.
(26) Mahmood, A.; Irfan, A.; Wang, J.-L. Machine Learning for Organic Photovoltaic Polymers: A Minireview. *Chin. J. Polym. Sci.* **2022**, *40*, 870−876.
(27) Greener, J. G.; Kandathil, S. M.; Moffat, L.; Jones, D. T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40−55.
(28) Magill, D. J.; Skvortsov, T. A. Depolymerase Predictor (Depp): A Machine Learning Tool for the Targeted Identification of Phage Depolymerases. *BMC Bioinf.* **2023**, *24*, 208.
(29) Mahmood, A.; Wang, J.-L. Machine Learning for High Performance Organic Solar Cells: Current Scenario and Future Prospects. *Energy Environ. Sci.* **2021**, *14*, 90−105.
(30) Kern, J.; Venkatram, S.; Banerjee, M.; Brettmann, B.; Ramprasad, R. Solvent Selection for Polymers Enabled by Generalized Chemical Fingerprinting and Machine Learning. *Phys. Chem. Chem. Phys.* **2022**, *24*, 26547−26555.
(31) Pandala, S. R. *Lazy Predict* 2021.
(32) Hussain, S.; Ali, S.; Shahzadi, S.; Rizzoli, C.; Shahid, M. Diorganotin(Iv) Complexes with Monohydrate Disodium Salt of Iminodiacetic Acid: Synthesis, Characterization, Crystal Structure and Biological Activities. *J. Chin. Chem. Soc.* **2015**, *62*, 793−802.
(33) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-Like Chemical Space Based on Gdb-17 Data. *J. Comput. -Aided Mol. Des.* **2013**, *27*, 675−679.
(34) Mahmood, A. Photovoltaic and Charge Transport Behavior of Diketopyrrolopyrrole Based Compounds with a−D−a−D−a Skeleton. *J. Cluster Sci.* **2019**, *30*, 1123−1130.
(35) Mahmood, A.; Khan, S. U.-D.; Rana, U. A.; Tahir, M. H. Red Shifting of Absorption Maxima of Phenothiazine Based Dyes by Incorporating Electron-Deficient Thiadiazole Derivatives as Π-Spacer. *Arabian J. Chem.* **2019**, *12*, 1447−1453.
(36) Khalid, M.; Khan, M. U.; Ahmed, S.; Shafiq, Z.; Alam, M. M.; Imran, M.; Braga, A. A. C.; Akram, M. S. Exploration of Promising Optical and Electronic Properties of (Non-Polymer) Small Donor Molecules for Organic Solar Cells. *Sci. Rep.* **2021**, *11*, No. 21540.
(37) Abdullah, M. I.; Janjua, M. R. S. A.; Nazar, M. F.; Mahmood, A. Quantum Chemical Designing of Efficient Tc4-Based Sensitizers by Modification of Auxiliary Donor and Π-Spacer. *Bull. Chem. Soc. Jpn.* **2013**, *86*, 1272−1281.
(38) Hussain, S.; Ali, S.; Shahzadi, S.; Shahid, M. Heterobimetallic Complexes Containing Sn(Iv) and Pd(Ii) with 4-(2-Hydroxyethyl)-Piperazine-1-Carbodithioic Acid: Synthesis, Characterization and Biological Activities. *Cogent Chem.* **2015**, *1*, No. 1029038.
(39) Hussain, S.; Ali, S.; Shahzadi, S.; Tahir, M. N.; Shahid, M. Synthesis, Characterization, Biological Activities, Crystal Structure and DNA Binding of Organotin(Iv) 5-Chlorosalicylates. *J. Coord. Chem.* **2015**, *68*, 2369−2387.
(40) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Chem.* **2009**, *1*, 8.
(41) Janjua, M. R. S. A.; Mahmood, A.; Ahmad, F. Solvent Effects on Nonlinear Optical Response of Certain Tetrammineruthenium(Ii) Complexes of Modified 1,10-Phenanthrolines. *Can. J. Chem.* **2013**, *91*, 1303−1309.
(42) Janjua, M. R.; Mahmood, A.; Nazar, M. F.; Yang, Z.; Pan, S. 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. Slovaca* **2014**, *61*, 382−390.

https://doi.org/10.1021/acsomega.3c05056
ACS Omega 2023, 8, 39408−39419