Journal of Materials Science and Chemical Engineering, 2024, 12, 1-11
https://www.scirp.org/journal/msce
ISSN Online: 2327-6053
ISSN Print: 2327-6045

Machine Learning Based Virtual Screening for Biodegradable Polyesters

Navya Nori

Milton High School, Milton, USA
Email: Navyanori6@gmail.com

How to cite this paper: Nori, N. (2024) Machine Learning Based Virtual Screening for Biodegradable Polyesters. *Journal of Materials Science and Chemical Engineering*, **12**, 1-11.
https://doi.org/10.4236/msce.2024.128001

Received: July 30, 2024
Accepted: August 19, 2024
Published: August 22, 2024

Abstract

Current biodegradation timelines show that polyesters take over 200 years to break down. A crucial component of several industries, polyesters are relied upon for materials development and thus require sustainable alternatives. Recent works in generative modeling have made it possible to produce large sets of chemical structures, but current molecular screening methods are expensive, not scalable, and are oversimplified. This work evaluates whether a molecule's biodegradability potential can be accurately predicted by training a model on recent experimental data. Additionally, three chemical descriptors were evaluated on the final molecules for their effects on biodegradability: molecular structure, bond types, and solubility. A Gradient Boosted Machine was trained on a dataset of 600 molecules and their binary labels on biodegradability. The classification model effectively captured the biodegradability property, yielding an Area Under the Receiver Operating Characteristics, AUROC, of 84% and an Area Under the Precision Recall Curve, or AUPRC, of 87%. Additionally, an existing amortized synthetic tree generation model, SynNet, validated each molecule by showing chemical synthesizability and producing simple and interpretable synthesis pathways. This approach of filtering by prediction and chemical rule interpretation is inexpensive, highly scalable and can capture the necessary complexity. Using this method, novel polyester candidates can be polymerized and produced into sustainable fabrics, reducing environmental stress from textile-reliant industries.

Keywords

Biodegradability, Molecular Generation, Virtual Screening

1. Introduction

Textile development for clothing is a primary and growing human need, and the

current demand is met by producing 110 million metric tons of industrial polyester per day [1]. The most commonly used polyester, which accounts for over half of all consumption, is polyethylene terephthalate (PET). The usage of fibers has doubled over the last 20 years due to the increased popularity of fast fashion companies [2]-[4]. These brands produce new styles and products weekly, optimizing for low cost rather than fabric biosafety. The rise of these industries has led to an increase in textile waste, amounting to 97 million tons in 2023 [5]. In anaerobic environments like landfills, these materials additionally produce 1.9 billion tons of CO_2 annually, causing air pollution and damage to human health [6]. Since polyesters take over 200 years to break down, these problems greatly worsen over time, increasing waste accumulation rates and reducing landfill space availability.

70% of all clothing is derived from polyester fibers, and the dyes used to color these pieces introduce toxicants and heavy metals into the water supply. These pollutants can adversely impact animal and human health when untreated wastewater enters local and residential water systems [5]. In the past 20 years, the global population has grown by 25%, leading to an increased demand for inexpensive clothing [7]. Thus, to support the population's growing needs while accounting for environmental challenges, sustainable alternatives to polyesters need to be developed at scale.

Over the past few years, researchers have developed innovative ways of generating molecules [8]. Initial approaches focused on atom-by-atom construction or substructure-based (ring or bond) assembly. These techniques perform best for small molecules, with significant declines in performance for large polymers [9]. In 2020, a method of molecular generation specifically tuned to large polymer generation, the Junction Tree Variational Auto-Encoder (JTVAE), was developed by Jin *et al.* [10]. This study developed a scalable method that generates polymers hierarchically, incorporating structural motifs, a key feature of large polymers, into the generation process. The generative model follows an encoder-decoder architecture in which the encoder takes a molecular input in the form of a graph and transforms it into a vector representation. The encoder captures both the coarse-grained motifs and the fine-grained atom connectivity. The decoder takes the embedding (lower dimensional representation of molecule) and constructs a new molecular graph step-by-step, adding one node or edge conditioned on the information learned from the encoder. The outputs of the decoder are thousands of novel graph representations of chemically valid molecules. However, these molecules are general and have different properties which need to be investigated and screened.

Previously, molecular screening has been done using two main methods: high-throughput screening (HTS) and knowledge-based filtering [11]. High-throughput screening is an experimental technique that rapidly evaluates the biochemical properties of millions of samples. Its high sensitivity and speed make it an optimal technique for simple and well-studied mechanisms. However, this method lacks support for more diverse systems and therefore is limited in its capacity for novel

polymer generation. This approach also has a high propensity for false positives and potentially lower hit rates due to incompatibility with other technology. This method has been used to effectively develop general high-volume datasets for further virtual and ML-based screening [12]. Fransen *et al.* conducted a biodegradation assay using the clear zone technique, where polymers were suspended under the action of bacteria *Psuedomonas lemoignei*. These observations were used to experimentally derive the effects of chemical structures on biodegradability across 600 polymers.

Virtual screening is a computational technique in which diverse sets of compounds are assessed for target properties. Current techniques in virtual and computational screening include ligand-based modeling and knowledge-based rules [13]. Ligand-based approaches predict molecular activity by mapping them to similar structures, which works well for well-studied molecules but tends to have a bias against novelty. Knowledge-based rules develop specific chemical criteria to select or exclude certain molecules. This targeted selection increases the likelihood of identifying optimal molecules and reliably eliminates clear outliers. Additionally, knowledge-based filtering can be tuned to specific application areas. Rules can be implemented specifically to filter molecules which are optimal for sustainability or drug discovery. However, these rules may be overly restrictive to the point of preemptively removing target molecules. For example, the Lipinski Rule of 5 contains specific quantitative boundaries for features such as molar mass, H donor and acceptor sites, and LogP. These restrictions lead to higher rates of false negatives because this form of filtering obeys the criteria regardless of biochemical complexity [14]. Additionally, rule-based filters typically cannot consider a broader application field, such as sustainability.

This work proposes a hybrid machine learning-based approach, in which JTVAE-generated polymers are filtered to biodegradable polyesters. For the filtering, a machine learning-based biodegradability classifier trained on Big Simplified Molecular Input Line Entry Systems, or BigSMILES, strings from HTS polyesters and polycarbonates, scores each generated molecule, capturing the biochemical complexity of biodegradability. The correlations of chemical rules with biodegradability are also computed to increase the interpretability of the final molecular design. The properties chosen to evaluate the structures from an atom, bonding, and environmental lens were molecular structure, bond types and interaction with water. For molecular structure, it is expected that aromatic rings will reduce biodegradability because of their high rigidity and absence of enzyme target groups, such as hydroxyls. Ester linkages are expected to increase biodegradability due to their high susceptibility to hydrolysis. Lastly, hydrophobicity will likely decrease biodegradability due to low solubility in water and other polar solvents.

2. Methods

2.1. Data Description and Preliminary Analysis

The methods used in this work can be split into three parts: polymer creation with

JTVAE generated molecules, biodegradable polyester filtering with a predictive model, and chemical and synthesizability analysis of the final molecules. The biodegradability predictor developed in this work uses gradient boosted trees, an AI/ML-model which retains performance even with small amounts of structured data.

We focus on molecules created by Jin *et al.*'s hierarchical encoder-decoder architecture for polymer graph generation. They used the polymer dataset with 86 K polymers from St. John *et al.* for training their model which was then tested for distributional statistics between original and generated compounds in addition to chemical validity and diversity [15]. In the current work, we generate 10,000 polymer molecules using model checkpoint 19 with the polymer vocabulary and other default settings for the library. Polyesters were then selected from the 10 k polymers by identifying molecules with a repeated ester linkage. Ester linkages form through the reaction between an organic alcohol and carboxylic acid, and a polyester is comprised of repeating units of this group.

Several chemical properties were computed for the molecular structures using the cheminformatics library RDKit, including LogP, molecular weight, heavy atom count, and bonding information [16]. These were computed to investigate the general characteristics of the polyesters, broadly understanding their molecular frameworks before applying sustainability-specific filtering.

2.2. Biodegradability Predictor

Biodegradability cannot be inferred from the above properties because structural motifs, interactions with solvents, and bonding information contribute to this vital characteristic. An AI/ML-based gradient boosted tree model was trained and tuned to quantify this property. The model was trained from a publicly available dataset containing molecules and binary labels showing their biodegradability from Fransen *et al.*'s work on the experimental discovery of biodegradable polymers.

The data was in the form of BigSMILES, string representations of molecular structure. The characters "{[][<]", "[<],[>]", and "[>][]}" were removed from the SMILES to facilitate encoding. These characters indicate the places of repeating units within a polymer, but do not impact the molecular formula. The model then implicitly learns that the subunit repeats. Molecules were created from the SMILES, and each molecule was verified to be kekulizable. In RDkit, kekulization involves fully expanding aromatic bonds, which contributes to standardization across molecules. Each molecule was then converted into a 128-bit Morgan fingerprint binary vectors, which leads to the modeling matrix having 128 columns and 549 rows.

The data was divided using a 70, 10, 20 split for training, validation, and testing, and the initial model was trained. This predictor computed a biodegradability score on each molecule, and the top ten highest-performing molecules were further analyzed. Several evaluation metrics were computed on the model, such as

Area Under Receiver Operating Characteristics curve (AUROC) and Area Under Precision Recall Curve (AUPRC). The validation dataset was used for early stopping when the binary cross entropy loss did not improve in 10 iterations. Hyperparameter tuning was performed on boosted tree parameters such as learning rate, maximum depth, minimum data in leaf, and regularization parameters lambda_l1 and lambda_l2.

2.3. Synthesizability Analysis

Gao *et al.*'s SynNet was used to create synthesis pathways to show if the polymers could be created in a laboratory using feasible reactions and purchasable compounds [17]. SynNet was chosen due to its choice of materials being preexisting monomers rather than generating unrealistic molecular structures.

SynNet uses a Markov decision process, which conditions on a target molecular embedding. The neural networks model synthetic trees according to reaction rules from a set possible space of reaction templates. The networks are trained on many pathways created from a database of compounds commercially available in the United States. This method was first validated by ensuring that the network could recover new molecules using conditional generation. Next, the method was validated through the identification of synthesizable structural analogs. Third, the molecules were validated through optimization for specific applications, such as drug discovery. In this work, SynNet was used to produce the SMILES strings of the monomers for each target polyester. The National Institute of Health's chemical Identifier Resolver was then used to find the compound name and make the results more interpretable [18].

3. Results and Discussion

3.1. Results for the Biodegradability Predictor

Figure 1 shows the Receiver Operating Characteristics curve of the biodegradability binary classification model for training, validation and test data splits. An AUROC of 83.59% on the holdout test dataset indicates that the model could effectively distinguish between biodegradable and nonbiodegradable molecules and capture the property very well. The black dotted line shows the minimum AUROC, only possible if the model learns no weights and operates completely randomly. Figure 2 shows the Precision Recall curve for the same model and data splits. AUPR of 87.24% on the test dataset indicates that the model is reliable and effective in identifying true positives, or labeled biodegradable polyesters.

3.2. Effect of Descriptors on Biodegradability

The chemically valid polymers generated by Jin *et al.*'s model were filtered as described previously and scored using the biodegradability model. The properties of the top ten polymers which had the highest score were further analyzed, and their synthesis pathways created.

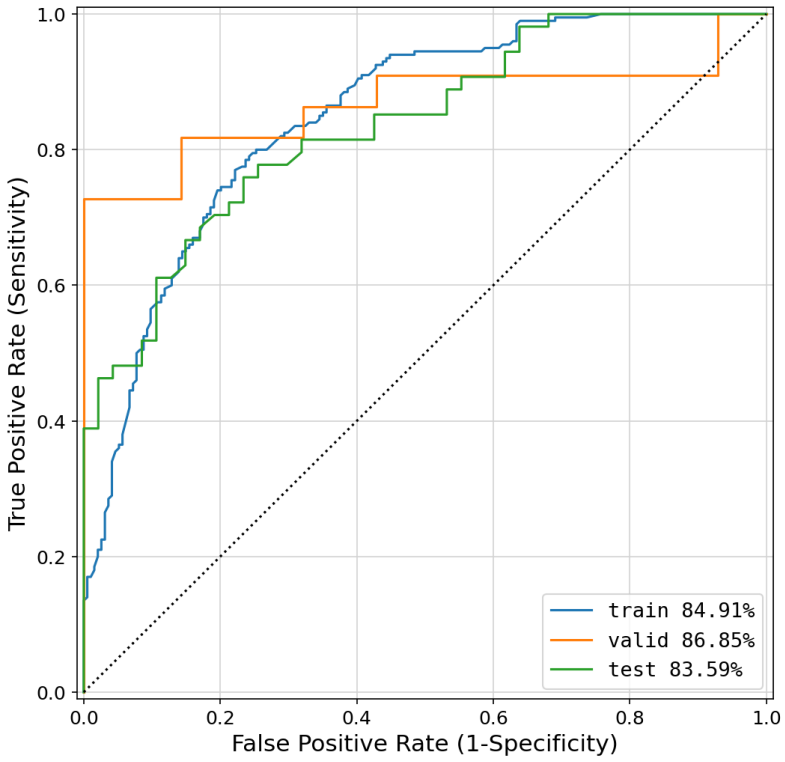

Figure 1. Receiver Operating Characteristics (ROC) Curves for train, validation and test data splits using the Biodegradability Prediction Model.

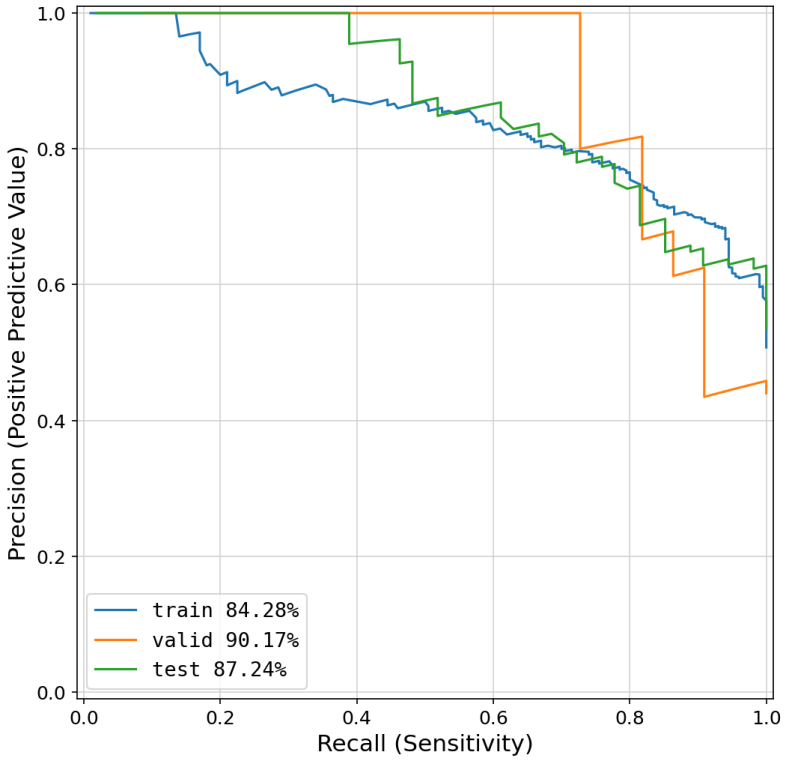

Figure 2. Precision Recall (PR) Curves for train, validation and test data splits using the Biodegradability Prediction Model.

The effects of various chemical properties on biodegradability are discussed below and are supported by the experimentally derived findings in Fransen *et al.'s* work. The most biodegradable molecules contained a carbon backbone of seven connections or lower; any molecules with greater than fifteen had largely inhibited biodegradability. This assertion is supported by the analysis of molecular weight as a chemical descriptor, which is negatively correlated with biodegradability. Also, polar heteroatoms (non-C or O atoms) contribute to biodegradability, likely supported by interactions with enzymes and their availability. **Figure 3** and **Figure 4** support these assertions about molecular weight. In **Figure 3**, a molecule with a score of greater than 0.8, the best performing molecule, contains a much shorter carbon backbone than the molecule in **Figure 4**, which scored very poorly.

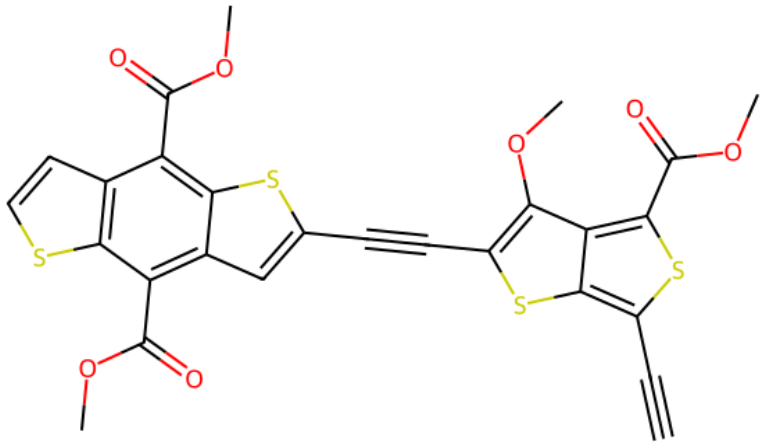

Figure 3. Visualization of a molecule scoring greater than 0.8.

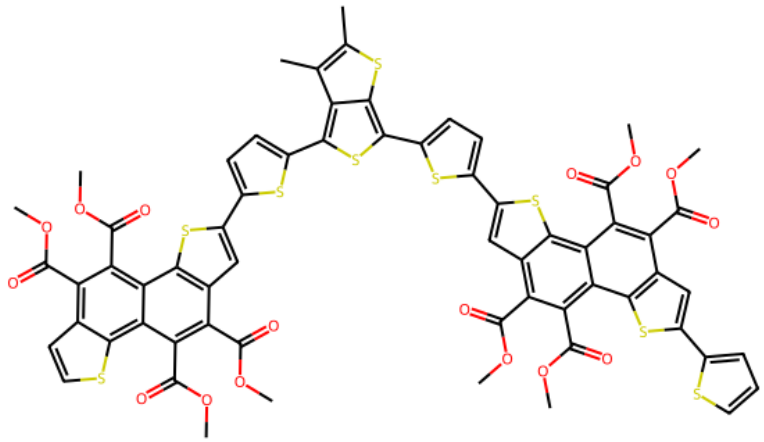

Figure 4. Visualization of a molecule scoring less than 0.8.

3.3. Effect of Molecular Structure on Biodegradability

The presence of aromatic rings had a weak negative correlation with biodegradability. This may be due to their higher rigidity, because of their alternating single and double bonds therefore decreasing biodegradability. In terms of material properties, the presence of these bonds may preserve features such as thermal and

mechanical strength, which are vital for polyesters to be used in textile production.

3.4. Effect of Bond Types on Biodegradability

For results on bond types, ester linkages were positively correlated with biodegradability, shown by a correlation coefficient of 0.062. This correlation may occur due to the following chemical properties. Ester bonds form between carbon and oxygen atoms, which have a substantial electronegativity difference. This results in a very polar bond, causing structures to be dissolvable in polar solvents and cleavable by hydrolytic enzymes. Thus, the presence of polar heteroatoms was also positively correlated with biodegradability, supported by the prevalence of sulfur atoms in Figure 3. Additionally, repeating C=O bonds specifically are commonly found in many types of organic matter. This prevalence has resulted in the evolution of metabolic degradation pathways of similar molecules across several microorganisms. Polyesters with these characteristics generally degrade in a wide variety of environments and conditions.

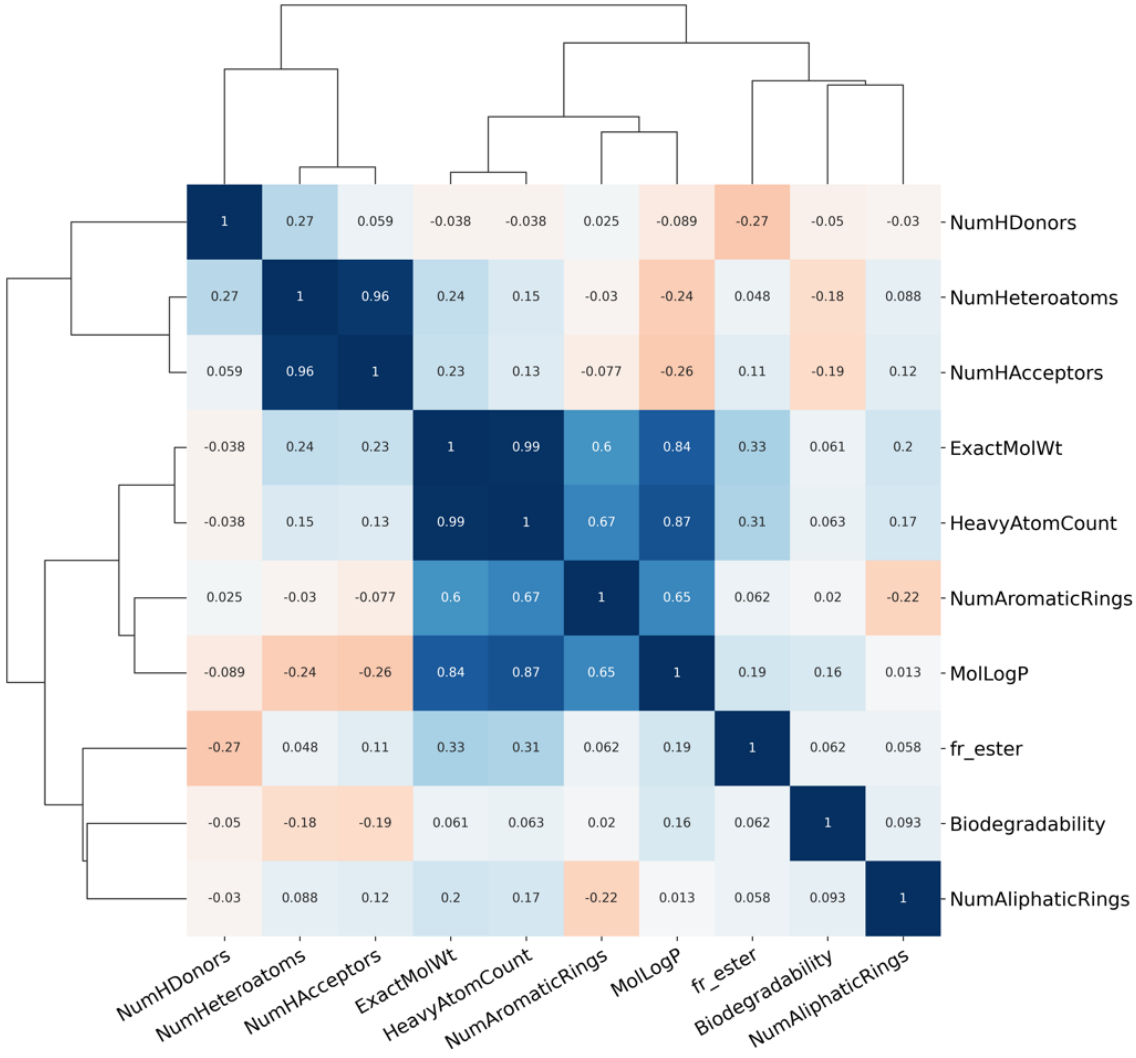

Figure 5. Correlation coefficients of chemical descriptors.

3.5. Effect of Solvent Interactions on Biodegradability

Finally, there was a weak positive correlation between hydrophobicity and biodegradation. This property was quantified using the partition coefficient, LogP. This measure is computed as a chemical descriptor in RDKit, and a positive LogP value demonstrates lipophilicity, while a negative value indicates a higher affinity for the aqueous phase. The correlation coefficient for LogP was 0.16, suggesting hydrophobicity was favored, likely due to a facilitated entrance through the cell membrane, increasing uptake and later degradation. Additionally, when polymerized, more hydrophobic molecules will retain qualities such as durability, breathability, and stain resistance, necessary properties in textiles. All analysis of chemical descriptors is supported by correlations calculated and shown in **Figure 5**.

3.6. SynNet Analysis

SynNet showed that the top high-scoring molecules are completely chemically synthesizable. The component parts having simplistic structures and less than ten monomers suggested the polyesters are accessible and not complex to generate. SynNet does not consider the commercial solvents and materials required to produce the polymers, only the basic components themselves. Thus, there may be additional materials and costs necessary to generate these new materials.

4. Conclusions

This study proposes a novel machine learning-based virtual screening method for biodegradable polyesters. To achieve this screening, a tree-based model captures the biodegradability property and scores a set of AI-generated molecules. The final molecules achieve the desired properties of biodegradability and synthesizability. Additionally, the method is reliable and representative, an improvement from the state of the art.

Molecular structure, bonding, and interactions with water were evaluated for their effect on biodegradability to aid in interpretability for the final molecules. The presence of aromatic rings was fairly neutral, suggesting that these rings in trace quantities may assist with final polymer properties. Presence of ester linkages positively correlates with biodegradability. Finally, hydrophobicity mildly increased biodegradability as shown by the partition coefficient.

Further analysis can include building prediction models that capture other properties which contribute to sustainability, such as solubility, and using them in conjunction with the biodegradability model. Additionally, the structures designed should undergo *in silico* simulations under various environmental conditions as well as *in vitro* testing. In the virtual screening space, generative models can be leveraged to produce novel compounds with desired properties, rather than filtering a set of general structures, posing several unique applications in the sustainability field.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Palacios-Mateo, C., van der Meer, Y. and Seide, G. (2021) Analysis of the Polyester Clothing Value Chain to Identify Key Intervention Points for Sustainability. *Environmental Sciences Europe*, **33**, Article No. 2. https://doi.org/10.1186/s12302-020-00447-x

[2] Bick, R., Halsey, E. and Ekenga, C.C. (2018) The Global Environmental Injustice of Fast Fashion. *Environmental Health*, **17**, Article No. 92. https://doi.org/10.1186/s12940-018-0433-7

[3] Sant'Ana, M.A. and Kovalechen, F. (2012) Evaluation of the Health Risks to Garment Workers in the City of Xambrê-Pr, Brazil. *Work*, **41**, 5647-5649. https://doi.org/10.3233/wor-2012-0906-5647

[4] Niinimäki, K., Peters, G., Dahlbo, H., Perry, P., Rissanen, T. and Gwilt, A. (2020) Author Correction: The Environmental Price of Fast Fashion. *Nature Reviews Earth & Environment*, **1**, 278. https://doi.org/10.1038/s43017-020-0054-x

[5] Khan, S. and Malik, A. (2013) Environmental and Health Effects of Textile Industry Wastewater. In: Malik, A., Grohmann, E. and Akhtar, R., Eds., *Environmental Deterioration and Human Health*, Springer, 55-71. https://doi.org/10.1007/978-94-007-7890-0_4

[6] Liu, Z., Deng, Z., Zhu, B., Ciais, P., Davis, S.J., Tan, J., *et al.* (2022) Global Patterns of Daily CO_2 Emissions Reductions in the First Year of Covid-19. *Nature Geoscience*, **15**, 615-620. https://doi.org/10.1038/s41561-022-00965-8

[7] Godfray, H.C.J. and Robinson, S. (2015) Contrasting Approaches to Projecting Long-Run Global Food Security. *Oxford Review of Economic Policy*, **31**, 26-44. https://doi.org/10.1093/oxrep/grv006

[8] Feng, W., Wang, L., Lin, Z., Zhu, Y., Wang, H., Dong, J., *et al.* (2024) Generation of 3D Molecules in Pockets via a Language Model. *Nature Machine Intelligence*, **6**, 62-73. https://doi.org/10.1038/s42256-023-00775-6

[9] Hu, W., *et al.* (2023) Deep Learning Methods for Small Molecule Drug Discovery: A Survey.

[10] Jin, W., Barzilay, R. and Jaakkola, T. (2020) Hierarchical Generation of Molecular Graphs Using Structural Motifs. https://arxiv.org/pdf/2002.03230

[11] Grygorenko, O.O., *et al.* (2020) Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience*, **23**, Article 101681. https://doi.org/10.1016/j.isci.2020.101681

[12] Fransen, K.A., Av-Ron, S.H.M., Buchanan, T.R., Walsh, D.J., Rota, D.T., Van Note, L., *et al.* (2023) High-Throughput Experimentation for Discovery of Biodegradable Polyesters. *Proceedings of the National Academy of Sciences*, **120**, e2220021120. https://doi.org/10.1073/pnas.2220021120

[13] Offutt, T.L., Swift, R.V. and Amaro, R.E. (2016) Enhancing Virtual Screening Performance of Protein Kinases with Molecular Dynamics Simulations. *Journal of Chemical Information and Modeling*, **56**, 1923-1935. https://doi.org/10.1021/acs.jcim.6b00261

[14] Hartung, I.V., Huck, B.R. and Crespo, A. (2023) Rules Were Made to Be Broken. *Nature Reviews Chemistry*, **7**, 3-4. https://doi.org/10.1038/s41570-022-00451-0

[15] St John, P.C., Phillips, C., Kemper, T.W., Wilson, A.N., Guan, Y., Crowley, M.F., *et al.* (2019) Message-Passing Neural Networks for High-Throughput Polymer Screening. *The Journal of Chemical Physics*, **150**, Article 234111. https://doi.org/10.1063/1.5099132

[16] Bento, A.P., Hersey, A., Félix, E., Landrum, G., Gaulton, A., Atkinson, F., *et al.* (2020) An Open Source Chemical Structure Curation Pipeline Using RDKit. *Journal of Cheminformatics*, **12**, Article No. 51. https://doi.org/10.1186/s13321-020-00456-1

[17] Gao, W., Mercado, R., and Coley, C.W. (2022) Amortized Tree Generation for Bottom-Up Synthesis Planning and Synthesizable Molecular Design.

[18] Cornell, A.P., Kim, S., Cuadros, J., Bucholtz, E.C., Pence, H.E., Potenzone, R., *et al.* (2024) IUPAC International Chemical Identifier (InChI)-Related Education and Training Materials through InChI Open Education Resource (OER). *Chemistry Teacher International*, **6**, 77-91. https://doi.org/10.1515/cti-2023-0009