DeepMyco - Dataset Generation for Dye Mycoremediation

Danika Gupta
The Harker School
San Jose, CA 95129
dan@gprof.com

Abstract

Textile dyes comprise 20% of global water pollution. Mycoremediation, a promising approach utilizing cheap, naturally growing fungi, has not seen scale production. While numerous studies indicate benefits, it is challenging to apply the specific learnings of each study to the combination of environmental factors present in a given physical site - a gap we believe machine learning can help fill if datasets become available. We propose an approach to drive machine learning research in mycoremediation by contributing a comprehensive dataset. We propose using advanced language models and vision transformers to extract and categorize experimental data from various research papers. This dataset will enable ML-driven innovation in matching fungi to specific dye types, optimizing remediation processes, and scaling up mycoremediation efforts effectively.

1 Introduction

Textile manufacturing is one of the world's greatest environmental polluters [1]. Textile dyes are responsible for 20% of global water pollution [2,3], with the relative damage growing daily to the finite freshwater on our planet. Furthermore, textile dyes in water have polluted agricultural areas and caused significant health damage to humans, animals, and plants [5]. Many techniques exist for processing textile dye effluent. However each method has positive and negative elements. For example, bioabsorption generates new forms of waste that need to be incinerated, utilized, or reprocessed [4]. A promising technique is mycoremediation, where natural fungal materials are used to break down the chemical structure of dyes into constituent components of CO2 and water. Mycoremediation has many potential advantages, including the ability to grow the substrate at low cost, generally understood positive interactions with soils, and the ability to degrade specific dye types. However, while substantial research exists on mycoremediation, few scale implementations exist [5]. While not conclusive, a recent review of patents in the field also indicates that there has not been a significant shift from research to production [5]. Known challenges include the fact that while many point solutions exist, each experiment is sufficiently unique, so it is challenging to generalize to a new case and feel confident about the specific method that should be used.

Figure 1 shows a simple example of the importance of process to decolorization efficacy. 150 mL of dye effluent was prepared by mixing 20 g of Rit Dye [17] in one liter of distilled water. One cup of trametes versicolor fungi was added and the combination was placed on a shaking table. After 2 weeks, the fungi were filtered out and another cup of fresh tramates versicolor was added. After 2 more weeks, the color of the resulting solution is measured via spectroscopy. A second experiment uses the same dye concentration, fungal mass, timeframe and agitation, but in this case placed all

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2024.

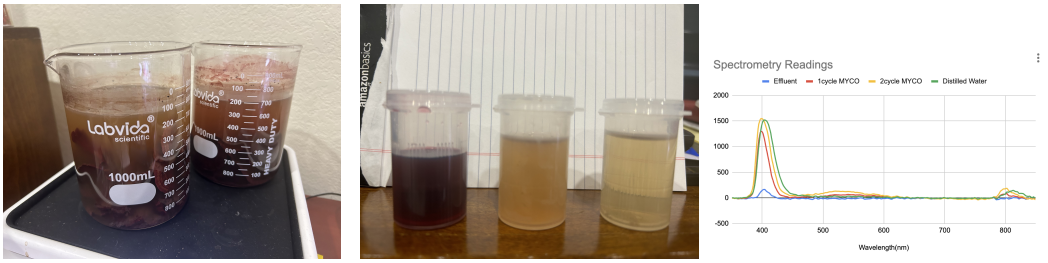

Figure 1: A Mycoremediation Experiment. The figures, left to right, show the experiment testbed, color change from start (leftmost), single cycle experiment (middle) and 2 cycle experiment (right). The final figure shows the spectrometry graph, indicating that while both experiments show value, the second approach gets much closer to distilled water

the fungi in the solution at the start and left it for four weeks. The decolorization levels are notably different.

This appears to be a problem to which machine learning can add value. The fundamental chemistry of mycoremediation, particularly for dyes, is known. However, the precise results are heavily dependent on environmental factors. Machine learning can discover the patterns within these relationships. There are more than 10,000 types of dye [7] and hundreds of strains of available fungi with mycoremediation potential [6], making it challenging to create simple models that can match fungi and dye. The challenge for applying machine learning is the lack of datasets. To our knowledge, no large-scale datasets exist for mycoremediation processes for dye treatment.

2 Methodology

Figure 2 describes our proposed methodology. We employ a web crawler to search for published research at the intersection of mycoremediation and dyes. Any publicly accessible PDF files are processed via a data processing pipeline to extract experiments contributed by each paper. The first step is to select whether the paper contributes unique experiments or is a review article. If it is the former, the PDF is processed in a number of ways (see Figure 2). with the goal of extracting one row of information for every unique experiment in the paper. By manually examining literature on dye mycoremediation, we have determined that the key factors affecting the performance of dye decolorization (besides the specific fungi and dye) are temperature, pH, agitation (shaking or stirring), timeframe, dye concentration, and fungal mass per unit of dye volume. The decolorization efficacy is often measured by color change spectroscopy and reported as a percentage improvement. Therefore, the pipeline attempts to extract each of these values for every unique experiment reported in every paper.

From an experimental standpoint, we intend to conduct an ablation study to explore the sensitivity of extraction effectiveness to different pipeline techniques. Figure 2 shows our planned study where each PDF is processed by using text extraction, segmented into pages with text-based retrieval augmented generation [9], page selection performed by a vision transformer, or fed directly into a large language model. The cross-sensitivity to the LLM itself will also be measured by testing several state-of-the-art Large Language Models (Lllama, GPT-4o, Gemini, and Claude). Effectiveness and correctness will be measured on a holdout set of research papers manually annotated for the correct experiments and then compared with the pipeline's extracted values. Measures will be reported of how many of the correct experiments were identified and missed, how many extraneous experiments were added, and the feature level correctness of all the correctly reported experiment rows. We intend to leverage a number of open source vision transformers and text processing methods in our study [18, 19, 20, 21, 23,24,25,26] as well as python processing tools [22].

3 Initial Feasibility Indicators

Our work to date has demonstrated several positive indicators for the feasibility of our approach. A crawler implemented to do a breadth-first search with deduplication, starting from a single recent mycoremediation paper, could access over 2000 relevant papers on the public internet in 24 hours, of which about 100 papers were found to be pertinent to our purpose. While this does not speak to

the total volume present and freely accessible, it is a positive indicator. Prior bibliometric research indicates that over 8000 research papers were published on textile dye treatment between 1990 and 2022 [16]. A sample analysis of the paper in [13] yields promise but also highlights the need for a comprehensive evaluation of text extraction approaches. [13] is a recent (2024) study of the mycoremediation efficacy of three fungal variants on five dye types. Each experiment generates seven results (one per day) for a total of 105 experiments. A straightforward query of GPT-4o delivered 15 experiments (the results of the 7-day outcome, all of which were correct in the columnar details), but could not extract the intermediate results, which were present in graphs in the paper. A tuned prompt requesting intermediate results returned 30 experiments, where some dye experiments were reported for multiple days and others for only one day. Of these, all but two were correct in all columnar details, providing preliminary indicators that our approach shows promise but detailed study and validation is needed to find the best extraction technique.

4 Pathway to Impact

We envision this dataset being used similarly to how [14,15] are used in the drug discovery process. These two datasets, both derived via NLP applied to public sources, has generated substantial innovation in their domain. We intend to use the methodology above and publish the method (and all generation code) with the dataset to demonstrate credibility. Furthermore, each fungal strain will be referenced via its MycoBank ID [1], ensuring accurate comparisons with other studies and enabling dataset users to link other resources such as the fungal DNA information provided by MycoBank. Via the MycoBank ID, dataset users can straightforwardly map each fungal strain to the enzymes within, which themselves are uniquely identifiable via the Enzyme Commission Number [8]. For example, Tramates Versicolor (Mycobank ID 281625 contains active enzymes Laccase (EC 1.10.3.2) which catalyzes the oxidation of phenolic compounds, Manganese Peroxidase (MnP, EC 1.11.1.13) which xidizes Mn(II) to Mn(III), which in turn oxidizes organic substrates, and Lignin Peroxidase (LiP, EC 1.11.1.14) which breaks down lignin by cleaving its non-phenolic structures. These fungal to enzyme mappings can be deterministically generated as needed by the study. Similarly, enzyme featurization for machine learning is an active research area [11,12]. We hope that state-of-the-art methods already discovered in other areas of enzyme engineering can be applied to the mycoremediation problem through the use of our dataset and that the dataset can contribute to new advances at the intersection of enzyme featurization, enzyme behavior prediction, and machine learning.

Code and dataset will be available at https://github.com/danikagupta/Deep-Myco as the project proceeds.

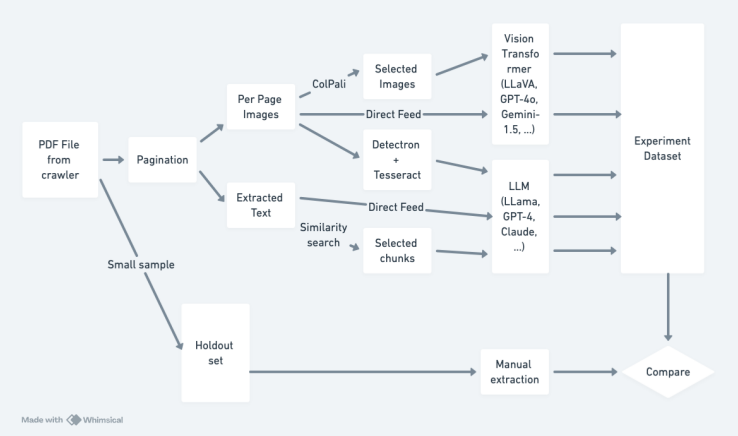

Figure 2: Study Methodology

References

[1] European Parliament. (2020) The Impact of Textile Production and Waste on the Environment - Infographics. Retrieved from https://www.europarl.europa.eu/topics/en/article/20201208STO93327/the-impact-of-textile-production-and-waste-on-the-environment-infographics.

[2] Cairns, R. (2023) One-fifth of water pollution comes from textile dyes. But a shellfish-inspired solution could clean it up. CNN. Retrieved from https://www.cnn.com/2023/06/15/world/textile-dyes-water-pollution-shellfish-solution-scn.

[3] Water Commission. (2023) Turning the Tide: A Report on Water Sustainability. Retrieved from https://watercommission.org/wp-content/uploads/2023/03/Turning-the-Tide-Report-Web.pdf.

[4] Tripathi, M., Singh, P., Singh, R., Bala, S., Pathak, N., Singh, S., Chauhan, R.S., Singh, P.K. (2023) Microbial biosorbent for remediation of dyes and heavy metals pollution: A green strategy for sustainable environment. Frontiers in Microbiology 14:1168954. DOI: 10.3389/fmicb.2023.1168954. PMCID: PMC10109241. PMID: 37077243

[5] Antón-Herrero, R., Chicca, I., García-Delgado, C., Crognale, S., Lelli, D., Gargarello, R.M., Herrero, J., Fischer, A., Thannberger, L., Eymar, E., Petruccioli, M., D'Annibale, A. (2024) Main Factors Determining the Scale-Up Effectiveness of Mycoremediation for the Decontamination of Aliphatic Hydrocarbons in Soil. Journal of Fungi.

[6]Harms, H., Schlosser, D., Wick, L.Y. (2011) Untapped potential: exploiting fungi in bioremediation of hazardous chemicals. Nature Reviews Microbiology, 9:177–192. DOI: 10.1038/nrmicro2519.

[7] Health the Planet. (n.d.) Textile Dyeing and Its Environmental Impact. Retrieved from https://healtheplanet.com/100-ways-to-heal-the-planet/textile-dyeing.

[8] Wikipedia. (n.d.) Enzyme Commission number. Wikipedia. Retrieved from https://en.wikipedia.org/wiki/Enzyme$_{C}ommission_{n}umber$

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D. (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. DOI: 10.48550/arXiv.2005.11401

[10] Crous, P.W., Gams, W., Stalpers, J.A., Robert, V., Stegehuis, G. (2004) MycoBank: an online initiative to launch mycology into the 21st century. *Studies in Mycology* **50**:19-22.

[11] Yang, J., Li, F.-Z., Arnold, F.H. (2024) Opportunities and Challenges for Machine Learning-Assisted Enzyme Engineering. ACS Central Science 10(2):141-150. DOI: 10.1021/acscentsci.3c01275.

[12] Salas-Nuñez, L.F., Barrera-Ocampo, A., Caicedo, P.A., Cortes, N., Osorio, E.H., Villegas-Torres, M.F., González Barrios, A.F. (2024) Machine Learning to Predict Enzyme–Substrate Interactions in Elucidation of Synthesis Pathways: A Review. Metabolites 14(3):154. DOI: 10.3390/metabo14030154. PMCID: PMC10972002. PMID: 38535315.

[13] Gugel, I., Summa, D., Costa, S., Manfredini, S., Vertuani, S., Marchetti, F., Tamburini, E. (2024) Mycoremediation of Synthetic Azo Dyes by White-Rot Fungi Grown on Dairy Waste: A Step toward Sustainable and Circular Bioeconomy. Fermentation, 10(2), 80. DOI: 10.3390/fermentation10020080

[14] Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2015 Oct 19. doi: 10.1093/nar/gkv1075

[15] Banda JM, Evans L, Vanguri RS, Tatonetti NP, Ryan PB, Shah NH. A curated and standardized adverse drug event resource to accelerate drug safety research. Scientific Data. 2016 May 10;3:160026. doi: 10.1038/sdata.2016.26

[16] Halepoto, H., Gong, T., Memon, H. (2024) Current status and research trends of textile wastewater treatments—A bibliometric-based study. Frontiers of Environmental Science.

[17] Rit Dye. (n.d.) Rit All-Purpose Dye. Retrieved from https://www.ritdye.com

[18] Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P. (2024) ColPali: Efficient Document Retrieval with Vision Language Models. arXiv preprint arXiv:2407.01449. DOI: 10.48550/arXiv.2407.01449.

[19] Facebook AI Research (FAIR). (2018) Detectron: Object Detection Platform. Retrieved from https://github.com/facebookresearch/detectron

[20] Smith, R. (2007) An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), IEEE, pp. 629-633. DOI: 10.1109/ICDAR.2007.4376991.

[21] Liu, Haotian, Linxi Fan, Chunyuan Li, Yong Jae Lee. "LLaVA: Large Language and Vision Assistant." 2023. Available: https://github.com/haotian-liu/LLaVA

[22] Maroš Kollár. (n.d.) PyPDF: A Python PDF Library. Retrieved from https://pypdf.readthedocs.io/.

[23] Anthropic. Claude, version 1.0, [AI language model]. Anthropic, 2023. Available: https://www.anthropic.com/claude

[24] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. "LLaMA: Open and Efficient Foundation Language Models." 2023. Available: https://arxiv.org/abs/2302.13971

[25] Hassabis, Demis, et al. "Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context." Google DeepMind, 2024. Available: https://arxiv.org/abs/2403.05530

[26] OpenAI. "GPT-4o." OpenAI Documentation, 2024. Available: https://www.openai.com/gpt-4o