JOURNAL OF
EMERGING INVESTIGATORS

Article

SOS-PVCase: A machine learning lignin peroxidase with polyvinyl chloride (PVC) degrading properties

Neel Ahuja[1], Emma S. Markowitz[2]
[1] Millburn High School, Millburn, New Jersey
[2] Harvard University, Cambridge, Massachusetts

SUMMARY

Plastic accumulating in landfills poses a threat to wildlife ecosystems and contributes to the production of harmful greenhouse gases. Polyvinyl chloride (PVC) accounts for 12% of plastic manufactured worldwide, and currently, 79% of post-consumer PVC ends up in landfills. Enzymatic degradation of plastic polymers into reusable monomers provides a green and scalable solution to this expanding problem. However, the application of PVC degrading peroxidases in real-world environments is impaired by their lack of stability and solubility. Fungal lignin peroxidase (E.C 1.11.1.14), an enzyme expressed by *Phanerochaete chrysosporium*, has previously been identified to have PVC degradation properties but, nevertheless, is hindered by the same constraints. In our study, we hypothesized that mutations in the primary structure of lignin peroxidase can improve the stability and solubility of the enzyme. To test our hypothesis, we utilized publicly available machine learning models to pinpoint stabilizing and solubilizing mutations of lignin peroxidase. The optimized mutant protein (SOS-PVCase: stable, optimized, soluble) contains five amino acid substitutions (A112I, A114I, S174K, E224M, L291R) which collectively improve stability by 18.6% and solubility by 10.2% compared to the wild-type enzyme, supporting our hypothesis that some mutations can enhance the protein's stability and solubility. This is important, as it allows PVC recycling pathways to be accelerated and made more economical. Furthermore, this suggests the potential of utilizing machine learning for the purpose of protein optimization.

INTRODUCTION

Polyvinyl chloride (PVC) has become ubiquitous in our modern lives, serving a multitude of purposes due to its versatility and durability. However, this convenience comes at a cost, as PVC poses a major environmental concern since it takes centuries to decompose naturally, releases harmful chemicals in the decomposition process, and contributes significantly to pollution (1). As a result, numerous techniques have been tested to depolymerize plastic polymers into recyclable monomers, which can then be reused in manufacturing plastic products. These breakdown techniques include enzymatic degradation, photodegradation, and mechanical degradation (2–4). Among the three, enzymatic degradation has long been hypothesized in the literature to be the most scalable and realistic process of accelerating plastic degradation pathways (5). Therefore, our study focuses only on metabolic degradation (using enzymes) as a method of breaking down PVC polymers.

Since 2005, a great deal of research has been centered around identifying specific enzymes capable of degrading polyethylene terephthalate (PET), a type of plastic that makes up many common consumer goods, such as water bottles and clothing (6–8). Moreover, further protein engineering studies have developed various computational algorithms to optimize these PET degrading proteins (PETases) for various properties, such as expanding the temperature and pH range they can function (9, 10). These approaches have been successful in the past, with some studies using these algorithms to identify specific missense mutations that improve both the stability and solubility of a protein (11–13). We simulated missense mutations, which are mutations in which a single DNA nucleotide change results in a different amino acid at a given position because of their ability to be generated using gene editing techniques. For example, some researchers have used gene editing techniques, such as CRISPR-Cas9, to create an optimized protein in a lab with missense mutations and verify the validity of machine learning optimization algorithms (14). One such study by Hongyuan et al. used machine learning models similar to the ones we used in our study to design an optimized PETase enzyme (15). They later synthesized this enzyme and, through various experiments, conclusively determined that it displayed improved properties compared to the wild-type enzyme. Similar to Hongyuan et al., our study used a process that optimized for protein stability and solubility (15). However, we used a novel meta-predictor variant approach, which combined two distinct machine learning models to make separate predictions about solubility and stability. This is expected to perform better than the single model used by Hongyuan et al., as the researchers in the study noted that some detail was lost in their predictions due to the singular model trying to optimize for both traits at once. To counter this, models in this study made isolated predictions, and then using statistics tests, we determined the mutations with the best combination of stability and solubility enhancement.

Furthermore, despite the increasing interest regarding the identification and mutagenic enhancement of PETases, little is known about PVC degrading proteins, called PVCases, and their potential applications. This is largely due to the fact that PVC polymers lack the hydrolyzable ester bond that is found in many other types of plastic, including PET, making its enzymatic degradation more challenging (16). As a result, to degrade PVC polymers into individual reusable vinyl chloride monomers, it is necessary to cleave the carbon bonds within

https://doi.org/10.59720/23-167

the polymer, which is a very energy-expensive reaction that only a few enzymes are capable of catalyzing (**Figure 1**). One PVCase of interest, *Phanerochaete chrysosporium*'s lignin peroxidase (PDB: 1LGA), shows promise, as Khatoon et al. observed a 31% PVC polymer weight loss using this enzyme over the course of four weeks (17). However, current PVCases, including *P. chrysosporium*'s lignin peroxidase, are hindered by their lack of stability and solubility compared to other plastic-degrading enzymes. Little research has gone into the optimization of the few identified PVCases, which is the gap that we aim to fill through our study.

Improving the stability and solubility of fungal lignin peroxidase would allow it to exist in a greater range of environments where its wild-type counterparts could not (18, 19). This is important for the application of fungal lignin peroxidase in landfills since the protein needs to remain stable in a variety of conditions to be useful in real-world scenarios (17). With the development of optimized PVCases being imperative to the faster recycling of PVC in landfills, identifying stabilizing and solubilizing mutations in the primary structure of lignin peroxidase is an important first step towards the large-scale degradation of harmful PVC plastic in a wide range of environments. Thus, before the mass production of fungal lignin peroxidase can begin, it is critical that we use accurate computational and machine learning models in our study to identify specific modifications that would enhance both the stability and solubility of the enzyme.

We used the Mutation Cutoff Scanning Matrix (mCSM), a novel supervised machine learning model trained using data from the Protein Data Bank (20), to predict the effect of missense mutations on the stability of fungal lignin peroxidase. This model was chosen over other protein-stabilizing machine learning methods in the literature due to the fact that it is a Convolutional Neural Network (CNN) and is trained on a very large and diverse data set (20). CNNs are inherently resistant to overfitting, which prevents bias in testing predictions and, thus, increases the accuracy of the overall model (21). Additionally, a data set with a great deal of variety also increases accuracy when it comes to testing predictions (22). This advanced architecture was the primary factor in the selection of this model over other protein stability prediction models in the literature, as they were trained on far smaller data sets and had higher rates of overfitting (23, 24). We also used Aggrescan3D to predict the effect each wild-type amino acid has on the solubility of a given protein (25). This model was chosen over others in the literature because it is a Recurrent Neural Network (RNN), and the specific hyperparameters used in the original report (including the number of layers and learning rate) work especially well for the task at hand (26). RNNs are extremely flexible, which is important for their successful application in enzyme solubility optimization, as enzymes come in a vast range of structures. These factors led to the selection of this model for our research study over the protein solubility prediction models created by Han et al. and Wang et al., as those algorithms had much slower learning rates and were not RNNs (27, 28).

Overall, we hypothesized that specific modifications to the primary structure of fungal lignin peroxidase could optimize both the stability and solubility of the enzyme. This goal was sufficiently accomplished, as the mutated enzyme, which we called Stable, Optimized, Soluble PVCase (SOS-PVCase), was predicted to display an 18.6% increase in Δ*G* (a metric used to determine stability) as well as have a 10.2% increase in solubility. This increase in protein stability and solubility is predicted to result in a greater range of environments where the enzyme can be used to break down harmful PVC in landfills.

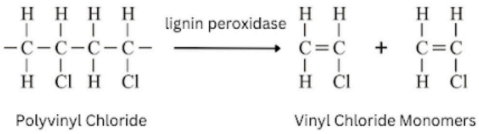

Figure 1: Polyvinyl chloride depolymerization reaction. Reaction scheme of PVC breakdown into vinyl chloride monomers catalyzed by fungal lignin peroxidase. The resulting vinyl chloride monomers can be reused into other PVC products.

RESULTS

mCSM Output

mCSM finds mutations stabilizing when the mutated amino is oriented closer to the geometric center of the protein than the wild-type amino acid at that position (making the protein more compact), while any mutation that increases this distance is considered destabilizing (20). Depending on the extent of the distance reduction, some mutations are more stabilizing than others. Using mCSM, we identified 712 stabilizing amino acid substitutions out of the 6,517 potential mutations for the 343 amino acids in lignin peroxidase. The unit for predicted change in stability by mCSM is the difference in Gibbs Free energy (Δ*G*) in kcal/mol between the wild-type and mutated protein. A decrease in distance between the amino acid and the geometric center of the protein is a process that requires an input of energy, making it a stabilizing mutation with a positive Δ*G* value. In contrast, an increase in distance between the amino acid and the geometric center of the protein is a process that releases energy, resulting in a destabilizing mutation with a negative Δ*G* value. Of the 712 stabilizing mutations, the ones with the highest stabilization rates were E168M, E168I, and E168L, with Δ*G*s of 3.034, 2.786, and 2.786, respectively (**Table 1**). Of all 6,517 potential mutations, the most destabilizing were F210G, F29D, and F322D, with Δ*G*s of -4.02, -4.051, and -4.069, respectively.

Wild-type Residue	Position	Mutant Residue	Mutation Predicted Δ*G* (kcal/mol)	Relative Stability Change (%)
E	168	M	3.034	12.036
E	168	I	2.786	11.052
E	168	L	2.786	11.052
D	264	M	2.473	9.810
D	238	M	2.457	9.747
E	168	P	2.36	9.362
E	168	V	2.36	9.362
D	238	I	2.274	9.021
D	238	L	2.274	9.021
D	264	I	2.006	7.958

Table 1: Effect of different mutations on protein stability. The top 10 most stabilizing mutations by mCSM ranked by the most beneficial impact on stability to least beneficial impact. The results were obtained by compiling all mutations into a .txt file and inputting the file into mCSM's server site, "Mutation List" mode, along with the PDB file of the protein. The single letters for residues correspond to the accepted nomenclature for those amino acids.

https://doi.org/10.59720/23-167

Aggrescan3D Output

Aggrescan3D finds mutations solubilizing when a hydrophobic amino acid near the outer regions of the protein is replaced with a hydrophilic amino acid (25). This is due to the concept of "like dissolves like," meaning a polar solvent, such as water, can only dissolve polar or hydrophilic substances. The amino acids on the outside of the protein actually come into contact with water, so having a hydrophilic amino acid near the outer regions of a protein instead of a hydrophobic amino acid would increase the protein's overall hydrophilicity and, thus, increase the overall solubility of the protein. Moreover, depending on how hydrophilic the mutated amino acid is and how much contact it has with the solvent (how close it is to the outer regions), mutations have different solubility scores.

Based on Aggrescan3D's output data about the lignin peroxidase wild-type, the most insoluble residues were determined. Subsequently, the amino acid substitutions that have a positive effect on the stability of the protein were inputted into Aggrescan3D to see if they also have a positive effect on the solubility. Afterwards, the solubility score of the wild-type amino acid was subtracted from the solubility score of the mutated amino acid at the same position to determine the true solubility score of a mutation. From this, the mutations that had a beneficial impact on both solubility and stability were determined. Of these mutations, the ones with the highest solubilizing rates were L291R, L291H, and F215E, with solubility improvement scores of 5.0846, 4.2509, and 3.402, respectively (**Table 2**). Of all mutations tested, the most insoluble ones were T76M, T327I, and T327L, with solubility deterioration scores of -1.160, -1.835, and -2.302, respectively. In total, there were 637 mutations that were stabilizing but resulted in a decrease in solubility between the wild-type and mutated protein.

Multiple Criteria Decision Analysis (MCDA) Output

After collecting this data from the optimization algorithms, we used the Multiple Criteria Decision Analysis (MCDA) statistical test to determine which mutations provide the best combination of stability and solubility improvement (**Table 3**). A mutation was considered to have a positive impact on these metrics if it leads to a positive relative change in stability and solubility, where the final value is the stability/solubility value of the mutated protein, and the initial value is the stability/solubility value of the wild-type protein. The five mutations with the highest MCDA scores were A112I, A114I, S174K, E224M, and L291R, which had MCDA scores of 1.304, 1.236, 1.218, 1.206, and 1.136, respectively. Collectively, they resulted in a 4.708 kcal/mol increase in the ΔG between the unfolded and folded state of fungal lignin peroxidase (stabilization) as well as an 11.951 increase in solubility score (**Table 3**), using Aggrescan3D's calculation metric. These overall scores were determined by summing the effect each of the five mutations had on protein stability and solubility. These values also correlate to an 18.6% increase in stability and a 10.2% increase in solubility when comparing the wild-type and mutated proteins.

DISCUSSION

We hypothesized that specific modifications could be made to the primary structure of fungal lignin peroxidase to improve the stability and solubility of the enzyme. Using mCSM and Aggrescan3D, we determined the effect of all mutations in the primary structure of fungal lignin peroxidase on protein stability and solubility. Of the 6,517 possible mutations in the primary structure of fungal lignin peroxidase (343 amino acids multiplied by the 19 possible substitutions), 712 had a positive impact on stability, and just 75 mutations were both stabilizing and solubilizing. After extracting the best mutation for each position from the data and ranking them based on the MCDA test results, we identified the top five mutations to optimize fungal lignin peroxidase for stability and solubility: A112I, A114I, S174K, E224M, and L291R (**Figure 2b**). The protein with these mutations was called SOS-PVCase.

In support of our hypothesis, SOS-PVCase (mutations A112I, A114I, S174K, E224M, and L291R) was predicted to be far more stable and soluble than the wild-type PVCase. Specifically, the mutated protein was predicted to display an 18.6% increase in ΔG as well as a 10.2% increase in solubility score. Moreover, the mutations collectively result in a +3 charge of the protein compared to the wild-type protein. This is likely to have made the protein more soluble because this increases the electrostatic attraction between solvent particles and the protein. For instance, a more positively charged protein can form stronger electrostatic attractions with the partially negatively charged oxygen atoms of water molecules, resulting in even greater dissolution of the protein by water solvent particles (30).

Wild-type Residue	Position	Mutant Residue	Mutation Solubility Improvement Score	Relative Solubility Change (%)
L	291	R	5.0846	4.354
L	291	H	4.2509	3.640
F	215	E	3.402	2.913
F	215	R	2.6166	2.240
C	317	M	2.5832	2.212
E	224	M	2.228	1.908
E	37	M	1.9331	1.655
S	53	L	1.8482	1.583
G	122	M	1.8012	1.542
S	53	M	1.7539	1.502

Table 2: Effect of different mutations on protein solubility. The top 10 most solubilizing mutations determined by Aggrescan3D ranked by most beneficial impact on solubility to least beneficial impact. Results were obtained by uploading the enzyme PDB file to the Aggrescan3D server site and customizing the options as described in the methodology.

Wild-type Residue	Position	Mutant Residue	% of Max Stability	% of Max Solubility	Overall Score
A	112	I	1.0000	0.3040	1.3040
A	112	L	1.0000	0.3040	1.3040
S	174	K	0.9322	0.3040	1.2363
A	114	I	0.9139	0.3040	1.2180
A	114	L	0.9139	0.3040	1.2180
L	291	R	0.2063	1.0000	1.2063
E	224	M	0.6988	0.4381	1.1369
D	185	M	0.7394	0.3040	1.0435
S	174	I	0.6852	0.3040	0.9893

Table 3: MCDA scores of various mutations. Beneficial mutations ranked by the statistically best combination of solubility and stability improvement to the worst combination. Rows highlighted in green indicate mutations included in the final optimized protein. The overall score is the sum of the percentage of maximum stability (in decimal form) and the percentage of maximum solubility (in decimal form).

https://doi.org/10.59720/23-167

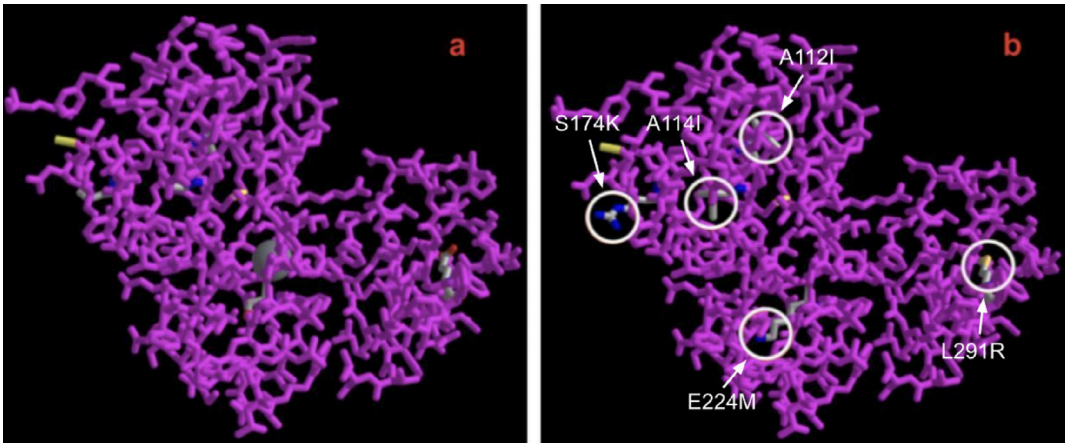

Figure 2: Simulated images of wild-type and mutated fungal lignin peroxidase. a) Wild-type fungal lignin peroxidase. Simulation generated by inputting PDB file of fungal lignin peroxidase into the NCBI protein generator (29). The amino acids in gray (including the gray sphere near the middle) are substituted in SOS-PVCase. **b)** Mutated fungal lignin peroxidase SOS-PVCase. The amino acid substitutions (A112I, A114I, S174K, E224M, L291R) are circled in white. Simulation generated by inputting PDB file of fungal lignin peroxidase into NCBI protein generator and applying mutations.

These increases in stability and solubility are sufficient to warrant a laboratory synthesis study to confirm the benefit of the mutations. In another study, synthesized proteins with structures optimized using XMutCompute, which is trained on data from the Protein Data Bank, similar to mCSM and Aggrescan3D, showed stability and solubility improvements consistent with the predicted values (15). Furthermore, mCSM and Aggrescan3D provide fast, computationally inexpensive, and accurate projections about the effect of various substitutions. This is important for the worldwide application of PVC-degrading enzymes in landfills, as a more soluble and stable alternative is critical for maintaining depolymerization functionality when exposed to a greater range of environmental conditions.

With that said one limitation of our study is our assumption that non-active site mutations will not inhibit the ability of the PVCase to catalyze the depolymerization reaction. While this is generally true, we cannot definitively prove that this is the case unless the mutated protein is synthesized and tested in a real-world experiment (31). If this assumption is proven incorrect, it is possible that when synthesizing SOS-PVCase, the enzyme is not able to complete the intended function of the degradation of PVC plastic. Furthermore, although mutation S174K, which was chosen to be included in SOS-PVCase, is positioned close to active site residues 176 and 177, it is unlikely that it will affect the protein's function because it is not actually in the active site. However, the only way to confirm that this mutation (and others) does not change the base protein function is to synthesize SOS-PVCase in a lab setting and conduct various degradation experiments.

Another limitation of our study is that it only included five mutations in the final optimized protein. It is possible that additional mutations that improve both stability and solubility can be induced without impeding the catalytic function of the protein. It is unlikely this is the case, as it has been found in literature that additional mutations, even if they are not at the active site, continuously increase the risk of impeding protein function, but a final conclusion can only be ascertained after synthesizing the mutant proteins (31). Additionally, to further validate our results, machine learning and computational algorithms that calculate protein stability and solubility using processes different from the ones used in our study, such as utilizing computer vision to identify unstable or insoluble areas in proteins, should be applied to wild-type fungal lignin peroxidase.

To conclude that the optimization mutations enhance PVC degradation as expected, we recommend that further investigations use gene editing techniques, such as CRISPR-Cas9, to synthesize the optimized SOS-PVCase. CRISPR-Cas9 is a relatively simple gene editing process, requiring only basic materials owned by most microbiology laboratories, such as injection microneedles, guide RNA, and restriction enzymes (32). In a series of controlled experiments, this optimized lignin peroxidase should be exposed to various temperature and pH ranges. Then, researchers should document dependent variables, including the exact temperature and pH of optimal protein function, which can be determined based on the production of vinyl chloride.

Notwithstanding the limitations, our study contributes to the current knowledge base of enzymatic PVC degradation, which severely lags behind those of other types of plastics. Furthermore, we used a novel meta-predictor variant approach that uses two separate machine learning models in combination with statistical tests (e.g. MCDA) to improve the stability and solubility of one PVC-degrading enzyme. In the future, other enzymes optimized using this approach can used for various applications, such as the military's initiative is preserving critical materials as well as getting rid of unwanted materials (33). Another potential application is in the Environmental Protection Agency's search for biological solutions to clean up oil spills, for which stable and soluble proteins would be essential (34). Additionally, the approach used in this study can be used to design enzymes that can assist the healthcare system in remediating the current medical waste issue, e.g., the accumulation of single-use latex gloves and masks (35). Overall, SOS-PVCase is versatile, as it has the potential to function in a greater range of environments (landfills), and it is novel since it has a new protein design.

https://doi.org/10.59720/23-167

MATERIALS AND METHODS

mCSM: A machine learning algorithm to optimize protein stability

$$\Delta G = RTln(r) \quad \text{[Eqn 1]}$$

mCSM calculates protein stability by predicting the Δ*G* (kcal/mol) of a mutation **(Equation 1)**. It does so by first calculating the distance, *r*, between the wild-type residue and the geometric center of the protein based on the three-dimensional structure. Then, the model is trained using a labeled data set where the features are amino acid substitutions, and the labels are the change in distance *r* compared to the wild-type residue. This distance *r* is vital to determining the effect of a missense mutation on protein stability. A decrease in the value (i.e., a mutation is closer to the geometric center) helps stabilize the protein, while an increase in *r* destabilizes the protein. mCSM then converts this *r* value to Δ*G* energy in kcal/mol using the formula above, where *R*=8.314J/K·mol is the ideal gas constant, *T* is the temperature in Kelvin (which in the mCSM simulation, is room temperature, i.e., 293K), and *r* is the distance of the amino acid from the geometric center.

Protein stability calculations were obtained using the mCSM web server. All sequence mutations were listed in a .txt file. The file was uploaded to the "Mutation List" page on the mCSM web server, along with the PDB file of the wild-type protein. The data was downloaded to a .csv file and analyzed in accordance with the results obtained from the Aggrescan3D output. mCSM predicts how mutations affect the stability of a protein by simulating all possible mutations at a given position. For example, at position one in the primary structure of fungal lignin peroxidase, which is originally the amino acid alanine, mCSM simulates all 19 other amino residues and returns the substitution that had the most beneficial impact on the stability of the protein as an output.

Mutations outputted by mCSM are identified in the following format: the single-letter abbreviation for the wild-type amino acid, followed by the residue position, followed by the single-letter abbreviation for the mutant amino acid. For example, E168M represents replacing the glutamic acid (E) at position 168 in the primary structure substituted with methionine (M). The 20 prominent amino acids existing in nature and tested in our study are as follows: arginine (R), histidine (H), lysine (K), aspartic acid (D), glutamic acid (E), serine (S), threonine (T), asparagine (N), glutamine (Q), cysteine (C), glycine (G), proline (P), alanine (A), valine (V), isoleucine (I), leucine (L), methionine (M), phenylalanine (F), tyrosine (Y), and tryptophan (W).

Aggrescan3D: A computational model to optimize protein solubility

Based on Aggrescan3D's output data of the wild-type lignin peroxidase, the residues that resulted in the greatest protein insolubility were determined (25). Aggrescan3D scored the solubility of wild-type residues by calculating aggregation or the abnormal association of proteins into larger structures, which tend to be insoluble, with lower scores indicating greater solubility (36). Proteins tend to aggregate when there are many hydrophobic amino acids near the outer regions of the three-dimensional structure of the protein (37). This causes the protein to be insoluble in various polar solvents and be ineffective for its intended purpose. Thus, increasing the solubility of lignin peroxidase through mutation enables the protein to be more effective at its function (in this case, the breakdown of PVC in landfills) for a greater duration of time without aggregation.

Subsequently, amino acid substitutions at these predicted insoluble positions, which mCSM also predicted to have a positive effect on the stability of the protein, were inputted into Aggrescan3D to determine if they also have a positive effect on the solubility. From this, data was collected about the solubilizing potential of various mutations by finding the difference between the wild-type solubility score and mutation solubility score.

Protein solubility calculations were obtained using the Aggrescan3D web server. The PDB file of the wild-type protein was uploaded to the web server along with the text "Chain A" (mutations were only induced on Chain A of the protein because the protein is only composed of a single polypeptide). The following options were selected: "Yes" for stability calculation, "Yes" for dynamic mode, "Yes" for mutated residues, "10A" for distance of aggregation analysis, and "Yes" for enhanced protein solubility.

Fungal lignin peroxidase active site

Information about the active sites of a protein is provided open source by the Protein Data Bank and the National Center for Biotechnology Information (38, 39). Primary structure sequence positions 43, 46, 47, 48, 66, 68, 70, 176, 177, 194, 196, 199, 201, and 238 are involved in the active site of *P. chrysosporium* lignin peroxidase (40). The identification of the active site of fungal lignin peroxidase is important because residues located at the active site cannot be substituted for another when trying to optimize the properties of a protein. This is because a small change introduced by genetic engineering in the active site of an enzyme can have a profound effect on the function of the protein, which is not the intention of our study (26). Thus, any mutation that was predicted to have a positive effect on the stability or solubility of fungal lignin peroxidase by mCSM and Aggrescan3D but was located at the active site, was disregarded as a potential mutation in the final optimized protein.

Multiple criteria decision analysis

A Multiple Criteria Decision Analysis (MCDA) statistics test (**Equation 2**) was used to determine the mutations that most improved the stability and solubility of fungal lignin peroxidase. Since the goal was to achieve both greater solubility and greater stability, the data was normalized by dividing each mutation's solubility or stability value by the maximum solubility or stability value achieved by any mutation. For example, if a given mutation had a solubility score of 0.1 and the maximum solubility score achieved by any mutation was 0.5, then the normalized solubility value of this mutation would be 0.2 (0.1/0.5). The same logic was applied to find the normalized stability of a mutation. Then, the overall "score" for any mutation was calculated by adding the normalized stability and solubility values. Finally, the mutations were ranked based on this overall score, and the mutations that provide the best combination of stability and solubility improvement were determined.

$$MCDA\ score = \Sigma\left(\frac{Value}{Max\ Value}\right) \quad \text{[Eqn 2]}$$

https://doi.org/10.59720/23-167

Determining SOS-PVCase

Many of the substitutions that were beneficial for both stability and solubility based on the MCDA test were mutants of the same wild-type position. For example, mutants of amino acid position 174 held the top two spots for the statistically best combination of solubility and stability optimization. The final mutant protein cannot have two mutations at the same position, and thus, only the mutation with the best score was selected to be included in the final optimized protein (SOS-PVCase) prediction. Furthermore, only five mutations were selected to be included in the final protein because, with any number greater than that, there is a very high risk of impeding the function of the protein, even with the absence of mutations in the active site (41).

ACKNOWLEDGEMENTS

We wish to thank Dr. Susan Arrigoni and Mr. Christopher Cook, advisors of the Science Research Program at Millburn High School, for their support and guidance during the research process.

Received: July 24, 2023
Accepted: November 21, 2023
Published: September 30, 2024

REFERNCES

1. Meng, Jun, et al. “Effects of Chemical and Natural Ageing on the Release of Potentially Toxic Metal Additives in Commercial PVC Microplastics.” *Chemosphere*, vol. 283, 1 Nov. 2021, pp. 131274, https://doi.org/10.1016/j.chemosphere.2021.131274.
2. Wilkes, Robert, et al. “Degradation and Metabolism of Synthetic Plastics and Associated Products by *Pseudomonas* Sp.: Capabilities and Challenges.” *Journal of Applied Microbiology*, vol. 123, no. 3, 31 May 2017, pp. 582–593, https://doi.org/10.1111/jam.13472.
3. Yousif, Emad, et al. “Photodegradation and Photostabilization of Polymers, Especially Polystyrene: Review.” *Chemistry and Materials Science*, vol. 2, no. 1, 23 Aug. 2013, https://doi.org/10.1186/2193-1801-2-398.
4. Corcoran, Patricia L. “Degradation of Microplastics in the Environment.” *Handbook of Microplastics in the Environment*, 14 Oct. 2020, pp. 1–12, https://doi.org/10.1007/978-3-030-10618-8_10-1.
5. Mohanan, Nisha, et al. “Microbial and Enzymatic Degradation of Synthetic Plastics.” *Microbiotechnology*, vol. 11, 26 Nov. 2020, https://doi.org/10.3389/fmicb.2020.580709.
6. Müller, Rolf-Joachim, et al. “Enzymatic Degradation of Poly(Ethylene Terephthalate): Rapid Hydrolyse Using a Hydrolase From *T. Fusca*.” *Macromolecular Rapid Communications*, vol. 26, no. 17, 5 Sept. 2005, pp. 1400–1405, https://doi.org/10.1002/marc.200500410.
7. Austin, Allen, et al. “Characterization and engineering of a plastic-degrading aromatic polyesterase.” *Biochemistry*, vol. 5, no. 3, 17 Apr. 2018, https://doi.org/10.1073/pnas.1718804115.
8. Palm, Reisky, et al. “Structure of the plastic-degrading *Ideonella sakaiensis* MHETase bound to a substrate.” *Nature Communications*, vol. 5, no. 3, 12 Apr. 2019, https://doi.org/10.1038/s41467-019-09326-3.
9. Wu, Cui, et al. “Deep learning-aided redesign of a hydrolase for near 100% PET depolymerization under industrially relevant conditions.” *Biological Sciences*, vol. 1, no. 1, 18 Jan. 2023, https://doi.org/10.21203/rs.3.rs-2465520/v1.
10. Gupta, Agrawal, et al. “Machine Learning-Based Enzyme Engineering of PETase for Improved Efficiency in Degrading Non-Biodegradable Plastic.” *Synthetic Biology*, vol. 1, no. 1, 12 Jan. 2022, https://doi.org/10.1101/2022.01.11.475766.
11. Han, Ning, et al. “Improving protein solubility and activity by introducing small peptide tags designed with machine learning models.” *Metabolic Engineering Communications*, vol. 11, no. 2, 1 Dec. 2020, https://doi.org/10.1016/j.mec.2020.e00138.
12. Sumida, Núñez-Franco, et al. “Improving Protein Expression, Stability, and Function with ProteinMPNN.” *Journal of the American Chemical Society*, vol. 3, no. 4, 9 Jan. 2024, https://doi.org/10.1021/jacs.3c10941.
13. Stanislav, Mazurenko, et al. “Predicting protein stability and solubility changes upon mutations: data perspective.” *ChemCatChem*, vol. 12, no. 22, 20 July 2020, https://doi.org/10.1002/cctc.202000933.
14. Thean, Chu, et al. “Machine learning-coupled combinatorial mutagenesis enables resource-efficient engineering of CRISPR-Cas9 genome editor activities.” *Nature Communications*, vol. 13, no.1, https://doi.org/10.1038/s41467-022-29874-5.
15. Lu, Hongyuan, et al. “Machine Learning-Aided Engineering of Hydrolases for PET Depolymerization.” *Nature*, vol. 604, no. 7907, 27 Apr. 2022, pp. 662–667, https://doi.org/10.1038/s41586-022-04599-z.
16. Zhang, Zhe, et al. “Polyvinyl Chloride Degradation by a Bacterium Isolated from the Gut of Insect Larvae.” *Nature Communications*, vol. 13, no. 1, 12 Sep. 2022, https://doi.org/10.1038/s41467-022-32903-y.
17. Khatoon, Nazia, et al. “Lignin Peroxidase Isoenzyme: A Novel Approach to Biodegrade the Toxic Synthetic Polymer Waste.” *Environmental Technology*, vol. 40, no. 11, 05 Jan. 2018, https://doi.org/10.1080/09593330.2017.1422550.
18. Fu, Hailong, et al. “Increasing Protein Stability: Importance of ΔC_p and the Denatured State.” *Protein Science*, vol. 19, no. 5, 25 Mar. 2010, pp. 1044–1052, https://doi.org/10.1002/pro.381.
19. Han, Xi, et al. “Improving Protein Solubility and Activity by Introducing Small Peptide Tags Designed with Machine Learning Models.” *Metabolic Engineering Communications*, vol. 11, 1 Dec. 2020, pp.138–140, https://doi.org/10.1016/j.mec.2020.e00138.
20. Pires, Douglas, et al. “mCSM: Predicting the Effects of Mutations in Proteins Using Graph-Based Signatures.” *Bioinformatics*, vol. 30, no. 3, 26 Nov. 2013, pp. 335–342, https://doi.org/10.1093/bioinformatics/btt691.
21. Alzubaidi, Zhang, et al. “Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions.” *Journal of Big Data*, vol. 8, no. 1, 31 Mar. 2021, https://doi.org/10.1186/s40537-021-00444-8.
22. McCloskey, Cox, et al. “Benefits of Using Blended Generative Adversarial Network Images to Augment Classification Model Training Data Sets.” *The Journal of Defense Modeling and Simulation*, vol. 3, no. 2, 29 Apr.

https://doi.org/10.59720/23-167

2023, https://doi.org/10.1177/15485129231170225.
23. Blaabjerg, Lasse et al. "Rapid protein stability prediction using deep learning representations." *Computational and Systems Biology*, vol. 2, no. 4, 15 May. 2023, https://doi.org/10.7554/eLife.82593.
24. Li, Yao, et al. "ProSTAGE: Predicting Effects of Mutations on Protein Stability by Using Protein Embeddings and Graph Convolutional Networks." *Journal of Chemical Information and Modeling*, vol. 4, no. 2, 2 Jan. 2024, https://doi.org/10.1021/acs.jcim.3c01697.
25. Kuriata, Iglesias, et al. "Aggrescan3D (A3D) 2.0: Prediction and Engineering of Protein Solubility." *Nucleic Acids Research*, vol. 47, no. 1, 3 May 2019, pp. 300–307, https://doi.org/10.1093/nar/gkz321.
26. Graves, Alex, et al. "Multi-Dimensional Recurrent Neural Networks." *Artificial Neural Neural Networks*, 1 Jan. 2007, pp. 549–558, https://doi.org/10.1007/978-3-540-74690-4_56.
27. Han, Ning, et al. "Improving protein solubility and activity by introducing small peptide tags designed with machine learning models." *Metabolic Engineering Communications*, vol. 11, no. 13, 1 Dec. 2020, https://doi.org/10.1016/j.mec.2020.e00138.
28. Wang, Zou, et al. "Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with DeepSoluE". *BMC Biology*, vol. 21, no. 12, 24 Jan. 2023, https://doi.org/10.1186/s12915-023-01510-8.
29. "ICn3D: Web-Based 3D Structure Viewer." *National Library of Medicine*. www.ncbi.nlm.nih.gov/Structure/icn3d/full.html?mmdbid=1LGA. Accessed 1 July 2023.
30. Salman, Muzammil, et al. "A Major Role for a Set of Non-Active Site Mutations in the Development of HIV-1 Protease Drug Resistance." *Biochemistry*, vol. 42, no. 3, 1 Jan. 2003, pp. 631–638, https://doi.org/10.1021/bi027019u.
31. Pelegrine, Gasparetto, et al. "Whey proteins solubility as function of temperature and pH." *Lebensmittel-Wissenschaft + Technologie/Food Science & Technology*, vol. 38, no. 1, https://doi.org/10.1016/j.lwt.2004.03.013.
32. Tong, Sheng, et al. "Engineered Materials for in Vivo Delivery of Genome-Editing Machinery." *Nature Reviews Materials*, vol. 4, no. 11, 4 Oct. 2019, pp. 726–737, https://doi.org/10.1038/s41578-019-0145-9.
33. Cremonesi, Paolo et al. "Enzymes as Tools for Conservation of Works of Art." *Journal of Cultural Heritage*, vol. 50, 1 July 2021, pp. 73–87, https://doi.org/doi.org/10.1016/j.culher.2021.06.005.
34. Rosenthal, André, et al. "Aqueous and Enzymatic Processes for Edible Oil Extraction." *Enzyme and Microbial Technology*, vol. 19, no. 6, 1 Nov. 1996, pp. 402–420, https://doi.org/10.1016/s0141-0229(96)80004-f.
35. Hudgins, Douglas. "Enzyme Diffusion and Cellulose Breakdown in the Bioremediation of Medical Waste." *vt.edu*, 7 Apr. 2009, https://doi.org/hdl.handle.net/10919/41928.
36. Weids, Ibstedt, et al. "Distinct Stress Conditions Result in Aggregation of Proteins with Similar Properties." *Scientific Reports*, vol. 6, no. 1, 18 Apr. 2016, https://doi.org/10.1038/srep24554.
37. Song, Jianxing. "Why Do Proteins Aggregate? "Intrinsically Insoluble Proteins" and "Dark Mediators" Revealed by Studies on "Insoluble Proteins" Solubilized in Pure Water." *National Library of Medicine*, vol. 2, 22 Mar. 2013, pp. 94–94, https://doi.org/10.12688/f1000research.2-94.v1.
38. *Research Collaboratory for Structural Bioinformatics Protein Data Bank*. National Institute of Health, 2024, www.rcsb.org/. Accessed 23 June 2024.
39. *National Center for Biotechnology Information*. National Library of Medicine, 2024, www.ncbi.nlm.nih.gov/. Accessed 23 June 2024.
40. "RCSB PDB - 1LGA: CRYSTALLOGRAPHIC REFINEMENT of LIGNIN PEROXIDASE at 2 ANGSTROMS." *Research Collaboratory for Structural Bioinformatics*, July 20, 2023, www.rcsb.org/structure/1LGA.
41. Alberts, Bruce. "Protein Function." *Molecular Biology of the Cell. 4th Edition.*, U.S. National Library of Medicine, 2002, www.ncbi.nlm.nih.gov/books/NBK26911/.