Un primer estudio estadístico de la Certificación en la UACM

Carlos E. Martínez-Rodríguez*

27 de noviembre de 2022

Índice

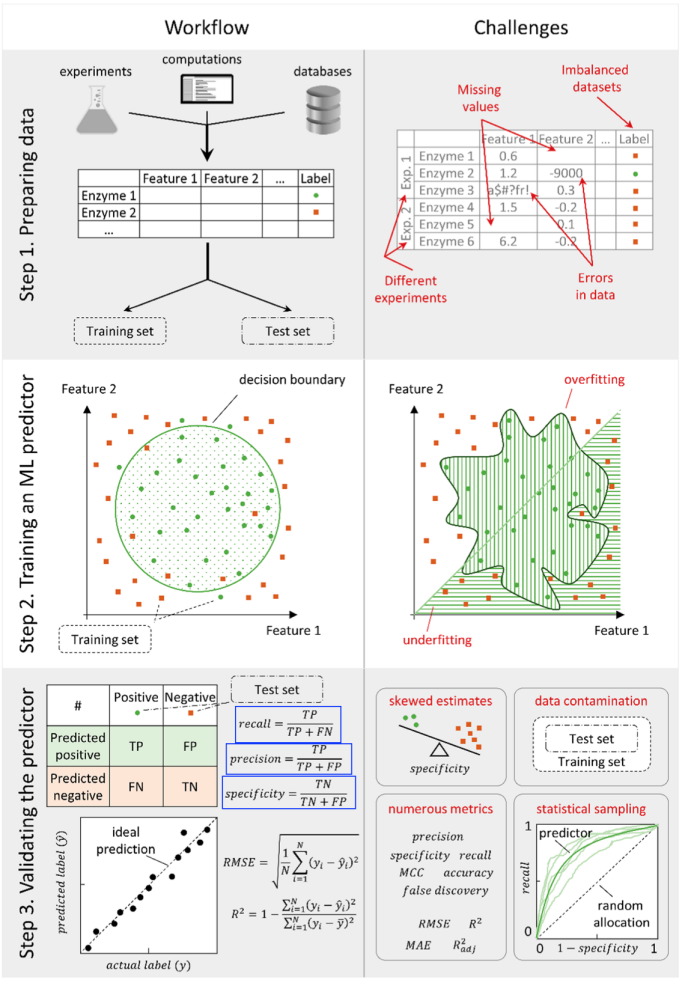
1. Introducción y antecedentes

1.1. Artículo 1: Machine Learning in Enzyme Engineering

Título: Machine Learning in Enzyme Engineering, Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky [1]

- Enzyme engineering is the process of customizing new biocatalysts with improved properties by altering their constituting sequences of amino acids.
- Multiple ML algorithms have already been applied to enzyme engineering. Some notable examples include random forests used to predict protein solubility [2], support vector machines [3, 4] and decision trees [5] to predict enzyme stability changes upon mutations, K-nearest-neighbor classifiers to predict enzyme function[6] and mechanisms,[7]

*Departamento de Estadística, Universidad Autónoma de la Ciudad de México (UACM). Correo electrónico: carlos.martinez@uacm.edu.mx



ht!

Figura 1: Schematic workflow of constructing an ML predictor and associated challenges.

and various scoring and clustering algorithms for rapid functional sequence annotation [8, 9]. The main attractiveness of ML in enzyme engineering stems from its generalizability: once it is trained on the known input, called a training set, an ML algorithm can potentially make predictions about new variants almost instantly.

- The aim of this Perspective is, therefore, to highlight recent advances in data collection and algorithm implementation for ML in enzyme engineering.

1.2. The essence of Machine Learning

La esencia de la mayoría de los algoritmos de Machine Learning (ML) es encontrar patrones en los datos disponibles, datos que consisten en varios descriptores o características, por ejemplo secuencias de encimas, sus estructuras secundarias y terciarias, substiruciones, etc. El nímero de características usualmente varían de decenas a miles lo que convierte el problema en uno de alta dimensión.

Los principales tipos de Machine Learning son: Aprendizaje Supervisado y Aprendizaja

No-Supervisado. En el aprensizaje no supervisado el objetivo es disminuir la alta dimensionalidad de los datos en uno de menor dimensión, o el de encontrar clústers en los datos. En el aprendizaje supervisado varias propiedades objetivo tales como actividad o estabilidad de enzimas, y el objetivo es diseñar un predictor que regrese etiquetas para datos no vistos considerando sus descriptores, utilizando el conjunto de datos etiquetado como datos de entrenamiento.

Nota 1 *Step 1: the data are usually turned into a table format and split into the training and test parts. Any errors, biases, or imbalances will be translated to the predictor's performance and, hence, must be accounted for. Step 2: the predictor is trained on the training data set. For example, a decision boundary is derived that allows classifying future input based on whether data points are inside or outside the boundary. This is a balancing act between two extremes: explaining noise rather than fundamental dependencies (overfitting) or failure to account for complex dependencies in the data (underfitting). Step 3: the performance of the predictor is evaluated based on the test data set. For example, true and false positives and negatives and the associated measures are calculated or the root mean square error (RMSE) is calculated for continuous labels. The random nature of the initial data split as well as data imbalances might skew the evaluation, and numerous metrics used for evaluation vary in their robustness to different data skews. Even partial inclusion of the test set at any stage of ML predictor training is called data contamination and usually invalidates the final evaluation.*

La etapa que más tiempo consume es la de recolección de datos y su preparación para alimentar el algoritmo de ML, entonces los datos son introducidos en el subconjunto de entrenamiento, el resultado se utiliza para meujorar los parámetros del predictor de ML, mientras que el segundo se utiliza para la evaluación.

Nota 2

- *En problemas de clasificación con etiquetas binarias o etiquetas con una cantidad finita de opciones, la evaluación usualmente se realiza por medio de la matriz de confusión: el número de verdaderos/falsos positivos y negativos.*

	Positivo	*Negativo*
Predecido Positivo	*TP*	*FP*
Predicido Negativo	*FN*	*TN*

- *Para problemas de regresión con etiquetas de valores continuos usualmente se calcula la raíz del error cuadrático medio*

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i - \hat{y}\right)^2} \tag{1}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}\right)^2}{\sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2} \tag{2}$$

En cualquiera de los dos casos la evaluación final se lleva a cabo en el conjunto de prueba, el cuál es esencial dado que el último objetivo es obtener el predictor más general en los datos no utilizados para entrenar el algoritmo.

Nota 3 *Las siguientes métricas se utilizan para medir el rendimiento de un modelo en función de su capacidad para predecir correctamente las clases de un conjunto de datos.*

- ***Recall (Recall o Sensibilidad):*** *Conocido como sensibilidad o tasa positiva real, mide la capacidad de un modelo para identificar correctamente todos los ejemplos positivos en un conjunto de datos. Se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos negativos:*

$$Recall = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Negativos} \tag{3}$$

  Un recall alto significa que el modelo es bueno para detectar los casos positivos, minimizando los falsos negativos. Es importante en situaciones donde los falsos negativos son costosos o críticos.

- *Precision (Precisión): La precisión mide la capacidad de un modelo para predecir correctamente los casos positivos entre todas las predicciones positivas que realiza. Se calcula como el número de verdaderos positivos dividido por la suma de verdaderos positivos y falsos positivos:*

$$Precision = \frac{Verdaderos\ Positivos}{Verdaderos\ Positivos + Falsos\ Positivos} \tag{4}$$

  Una alta precisión significa que el modelo tiene una baja tasa de falsos positivos, es decir, que cuando predice una clase como positiva, es probable que sea correcta. La precisión es importante en situaciones en las que los falsos positivos son costosos o no deseados.

- *Specificity (Especificidad): La especificidad mide la capacidad de un modelo para predecir correctamente los casos negativos entre todas las predicciones negativas que realiza. También se conoce como tasa negativa real. Se calcula como el número de verdaderos negativos dividido por la suma de verdaderos negativos y falsos positivos:*

$$Specificity = \frac{Verdaderos\ Negativos}{Verdaderos\ Negativos + Falsos\ Positivos} \tag{5}$$

  Una alta especificidad indica que el modelo es bueno para identificar correctamente los casos negativos, minimizando los falsos positivos. Esto es importante en situaciones en las que los falsos positivos son costosos o problemáticos.

Estas métricas proporcionan una forma más completa de evaluar el rendimiento de un modelo de clasificación que simplemente mirar la precisión general.

En la ingeniería de proteínas, las similitudes en secuencias en ambos subconjuntos de datos deben ser tenidas en cuenta. Si alguna familia de proteínas está sobre representada en el conjunto de prueba, el predictor resultante puede resultar sesgado hacia la identificación de patrones válidos solamente para esta familia. Si algunas secuencias en el conjunto de

prueba son muy cercanas al conjunto de entrenamiento, la evaluación final de desempeño dara resultados sobre optimistas.

En el paso 2 de entrenamiento, es posible ajustar el predictor o seleccionar de entre varios predictores, usualmente por medio de validación $k - fold$. En este caso los datos de entrenamiento se subdividen en K subconjuntos y el flujo de trabajo se repite K veces, con cada uno de ellos utilizaods para la evaluación de los $K - 1$ subconjuntos utilizados para entrenar. El reto principal en el paso 2 para cualquiern entrenamiento tipo ML supervisado es evitar el subajuste de los datos (sesgo alto) y el sobre ajuste (varianza grande).

La **subestimación** ocurre cuando un predictor falla en encontrar patrones incluso en los datos de entrenamiento (cuando un modelo lineal simple se utiliza para explicar dependencia dependencias no lineales en los datos). El **sobreajuste** ocurre cuando el desempeño de un predictor disminuye notablemente en los datos de prueba en comparación con los datos de prueba, debido al aprendizaje de demasiado detalle y ruido, en lugar de identificar patrones generales. Tanto el subajuste como el sobreajuste pueden ser debido a la insuficiente calidad de los datos: ruido excesivo, características faltantes o irrelevantes, sesgo en los datos, o datos dispersos. También pueden ocurrir como consecuencia de una pobre aplicación del algoritmo: excesiva o insuficiente flexibilidad en la selección de los parámetros, protocolo de entrenamiento inapropiado, o contaminación de los datos de entrenamiento con el conjunto de datos de prueba.

1.3. Bases de datos relevantes a Ingeniería de Enzima

1.3.1. The State of the Art in Data Accumulation

Debido a que los algoritmos de ML se basan en los datos, la importancia de la calidad de los mismos utilizados para entrenamiento no puede ser subestimada.

Ejemplos de conjuntos de bases de datos, utilizadas en la ingeniería de enzimas, son secuencias de proteínas y estructuras de proteínas. La estabilidad y solubilidad de las proteínas son dos cualidades que han sido medidas por varias décadas y hasta la fecha. Tareas más desafiante es la anotar las propiedades catalíticas de las enzimas debido a la abundancia de tipos de reacciones, mecanismos, cofactores, amplios rangos de especificidades de sustratos, enancionselectividades y promiscuidades.

Las enzimas son catalizadores biológicos que facilitan una amplia variedad de reacciones químicas en los organismos vivos. Algunas razones por las cuales la anotación de sus propiedades catalíticas es complicada son:

- **Tipos de reacciones diversos:** Las enzimas pueden catalizar una amplia gama de reacciones químicas, incluyendo reacciones de oxidación-reducción, hidrólisis, condensación, isomerización y más. Cada tipo de reacción involucra mecanismos químicos y sustratos diferentes.

- **Mecanismos:** Incluso dentro de un solo tipo de reacción, las enzimas pueden emplear múltiples mecanismos. Comprender el mecanismo específico utilizado por una enzima requiere un conocimiento detallado de la estructura de la enzima y de su sitio activo.

- **Cofactores:** Muchas enzimas requieren cofactores, como iones metálicos o coenzimas, para catalizar reacciones de manera efectiva. Identificar los cofactores necesarios para cada enzima es esencial para la anotación.
- **Condiciones de reacción:** La actividad enzimática puede depender en gran medida de las condiciones ambientales, incluyendo la temperatura, el pH y la fuerza iónica. La anotación de las condiciones óptimas para la actividad enzimática es crucial.
- **Especificidades de sustratos:** Las enzimas pueden ser altamente específicas para ciertos sustratos, reconociéndolos con alta afinidad, mientras que otras son más promiscuas y pueden unirse a una variedad de sustratos. La caracterización de la especificidad de sustratos es compleja.
- **Enantioselectividades:** Algunas enzimas pueden discriminar entre enantiómeros (isómeros de imagen especular) de una molécula, catalizando reacciones con alta selectividad por un enantiómero. La anotación de esta propiedad implica comprender la estereoquímica.
- **Promiscuidades:** Las enzimas pueden exhibir actividades promiscuas, catalizando reacciones diferentes a su función principal. Detectar y caracterizar tales promiscuidades es complicado.

1.3.2. Current Challenges Related to Databases

Si la dependencia que se busca no se encuentra en los datos disponibles, no importa la cantidad de nuevos datos ayudarán a mejorar la calidad del predictor de ML. En el caso de la ingeniería de enzimas se espera que las funciones enzimáticas estén codificadas en las secuencias y así depender en las propiedades físico-químicas de los aminoácidos, de aquí que la cantidad y la calidad de los datos en las bases de datos sean de importancia para diseñar un predictor de ML.

La falta de estándares en los reportes resulta en pérdida de información o valores erróneos para algunos descriptores. A esto hay que agregar la falta de protocolos robustos en los análisis de datos, como por ejemplo aquellos utilizados para ajuste de curvas para determinar las temperaturas de fusión o constantes cinéticas. Otro factor es que los recientes avances vuelven obsoletos resultados previos. La curación manual ayuda mejorar la calidad de los datos, sin embargo no se encuentra excenta de errores de anotación de las funciones de las proteínas y errores de propagación a partir de resultados previamente refutados.

Este tipo de procedimientos, verificación manual, puede incluir la limpieza o formato de datos para que sean amigables con ML. Uno de los principios más populares es **FAIR**, por sus siglas en inglés, Findable, Accesible, Interoperables y Reutilizables, debería de facilitar a las computadoras para que de manera automática pueda encontrar y utilizar los datos. Para las enzimas la guía estándar para reportar datos de enzimas (STRENDA) debería de aumentar la calidad de los datos, especialmente en bases de datos (bdd) heterogéneas recopiladas de diversas fuentes.

El desarrollo de nuevos predictores de ML ha incrementado considerablemente la demanda de mejora de las bdd existentes, así como la generación de nuevas bdd uniformes y representativas de mayor calidad.

Existen varias nuevas técnicas emergentes tales como

i Secuenciación de nueva generación.

ii Clasificación de células activadas por fluorescencia.

iii Exploración mutacional profunda, y

iv Microfluidos.

1.3.3. Emerging Methods for High-Throughput Data Collection

Avances tecnológicos hacia la miniaturización, automatización y paralelización han generado tecnologías eficientes de nuevos métodos de investigación experimental con capacidades incomparables. Secuenciación de nueva generación (NGS) ha revolucionado la investigación genómica, habilitado el acceso a datos moleculares fundamentales y revelado firmas genómicas y transcriptomicas [10, 11].

La capacidad de secuencias en el rango de gigabases por ejecución del instrumento permite secuencias el genoma humano en su totalidad en tan sólo un día.

Múltiples instrumentos comerciales de segunda generación disponibles ofrecen mayor capacidad y precisión. Métodos de tercera generación recientemente introducidos (lectura larga) que emplean secuenciación en tiempo real de moléculas [12] o secuenciación de nanoporos [13] resuelve las limitaciones de lecturas cortas, como el sesgo de GC o mapear elementos repetitivos

Mientas el avance de la tecnología de secuenciación proporciona una gran cantidad de secuencias de datos, para la mayoría de estas entradas, las anotaciones estructurales y anotaciones funcionales aún están perdidas.

Como siguiente paso, se está centrando en el desarrollo de nuevos métodos experimentales efectivos para recopilar información funcional y estructural.

La clasificación de células activadas por fluorescencia (FACS) es una tecnología ampliamente disponible que permite el cribado de hasta 108 variantes de enzimas por día. La FACS requiere que los sustratos fluorogénicos estén atrapados dentro o en la superficie de la célula para vincular el genotipo y el fenotipo. Alternativamente, se utiliza la clasificación de enzimas encapsuladas junto con su ADN codificador y un sustrato fluorogénico en perlas de hidrogel.

Cuando se combinan con la secuenciación de próxima generación, los ensayos de alto rendimiento representan una estrategia poderosa para analizar de manera integral las relaciones entre secuencia y función en las enzimas [14, 15]. Este enfoque, llamado escaneo mutacional profundo (DMS por sus siglas en inglés), vincula el genotipo con el fenotipo sin necesidad de procesos laboriosos que involucren la purificación y caracterización de proteínas. Durante el proceso, se sintetiza una gran biblioteca de secuencias mutantes, seguida de la selección de fenotipos expresados. Luego, la secuenciación de la biblioteca antes y después de la selección cuantifica la aptitud de cada mutante. De esta manera, el DMS proporciona un método rápido y sencillo para inferir los determinantes de la secuencia de la estabilidad y la función de las proteínas[14, 16, 17]. El DMS se ha utilizado como una estrategia experimental alternativa para la determinación de la estructura de las proteínas.

1.4. MACHINE LEARNING APPLICATIONS TO ENZYME ENGINEERING

Despite being a relatively new field of study, machine Learning for enzyme engineering has already been applied for several challenging predictions. First consider predictors aimed at elucidating the structurefunction relationships crucial for enzymes on both sides:

- predicting the structure for a known sequence, and
- predicting the catalytic activity or substrate specificity for a known sequence/structure.
- solubility and stability, from the point of view of amino acid substitutions,

The protein structure prediction is arguably one of the longest- standing challenges in biochemistry, as the number of resolved structures is dramatically lagging behind the number of known sequences. Over 145000 structures have been released in the Protein Data Bank, but this is still nowhere near over 215 million publicly available protein sequences[18]. Nevertheless, even despite a relatively small data set size in comparison to millions of data points usually available for this method, deep neural networks showed most the notable results in the latest biennial assessment of protein structure prediction methods, CASP13.

The AlphaFold network was trained on the PDB entries to predict the distances between C-beta atoms of residues using multiple sequence alignments[19] and received the highest score at the competition. Out of 124 targets, around two-thirds of AlphaFold predictions had a GDT_TS score above 50, which is indicative of a topologically correct structure [20].

Despite showing a tremendous improvement on the CASP12 results, this still indicates enough room for further improvement of protein structure predictors. Apart from predicting protein structures, predicting catalytic activities is another active field of research currently. Computational methods for the protein function prediction range from sequence- to structure-based and from gene- to genome- and interactome-based[21]. Several initiatives similar to the CASP competition have already been proposed to address the functional annotation of enzymes, namely Enzyme Function Initiative (EFI), the Computational Bridges to Experiments initiative (COMBREX), and the Critical Assessment of Function Annotation community-driven experiment (CAFA). Certain successful attempts to apply ML to assign enzyme EC numbers using predicted 3D structures [22] or exploiting sequence similarities [23] have already been made. Recently, deep learning was also applied to predict EC numbers on the basis of a protein sequence using both sequence-length-dependent features, such as raw sequence one-hot encoding, and sequence-length-independent features, such as functional domain encoding [24]. The former type of features introduced nonuniformity in feature dimensionality, and the authors presented a framework to perform simultaneously dimensionality uniformization, feature selection, and classification model training. As the validation for their predictor, activities of three isoforms of glutaminase and five isoforms of Aurora kinases B were predicted in good correspondence with the experimental data available. Thus, the large data sets of enzyme structures and activities accumulated to date already allow using deep learning in the engineering of catalytic activity. Nevertheless, the problems with the data sets mentioned earlier are aggravated in the case of recording enzyme activity profiles due to both complex nomenclature and the abundance of possible

mechanisms. A more precise functional prediction is possible by restricting ML training to a particular family of enzymes, which comes at the cost of much smaller data sets available for training. This problem may be tackled by applying high-throughput data collection methods mentioned before. The authors of the recently released GT-predict [25] selected for their analysis the glycosyltransferase superfamily 1, a group of enzymes with highly diverse substrates. This diversity, combined with the high scaffold conservation, increases the importance of subtle background mutations for the chemical function. Data from the label-free mass spectroscopy-based assay of 91 substrates and 54 enzymes derived from the plant Arabidopsis thaliana were used for functional prediction. The authors trained sequence-based decision trees, systematically varying combinations of physicochemical properties, e.g. log P, molecular area, and number/type of nucleophilic groups, and structural information, e.g. scaffold type and functional groups. The resulting predictor was successfully tested on four individually selected gene sequences as well as two complete families of enzymes from four different organisms, which highlights the tremendous potential of training ML predictors on the newly acquired data from high-throughput data collection methods. However, caution must be taken when extrapolating the results of this study to other families. It is yet to be seen if a strong predictor for one family will perform well when it is retrained on the data for another family. Predictors of protein solubility usually exploit the eSol database (Table 1) for the entire ensemble of Escherichia coli proteins [26] In their recent paper [29] Han and coauthors considered seven different binary and continuous ML algorithms: logistic regression, decision tree, support vector machines, Naive Bayes, conditional random forest, XGboost, and artificial neural networks. The support vector machine

The authors attempted to use generative adversarial networks to synthesize more data. This is a pair of two neural networks competing against each other: one learns to generate artificial examples and the other to distinguish them from real data. However, due to data scarcity, no independent test set was used to evaluate the resulting predictor, implying there is a strong demand for more abundant data sets of protein solubility. Moreover, a modest best-achieved R2 value of around 0.4 indicates that there is still ample room for designing a more reliable continuous predictor of solubility scores. Another point of view on protein solubility prediction is studying the effects of individual mutations. The recent successes in the application of deep mutational scanning to collect the data on protein solubility changes upon mutations[27] are likely to promote the development of more sophisticated ML-based protein solubility predictors in the nearest future. Predicting the effects of amino acid substitutions is not only limited to solubility: stability, substrate specificity, catalytic activity, and enantioselectivity can also be targeted if sufficient data are available. Protein stability predictors are perhaps those with the most abundant data sets of this type available for ML training (Table 1). The recently released PON-tstab34 stands out due to the impressive work the authors undertook to identify major issues with the widely used ProTherm database. The authors also presented a random forest classifier trained using 1106 features from the following groups: experimental conditions, conservation and coevolution scores for mutated positions, amino acid substitutions and their physicochemical properties, neighborhood features for 11 positions before and after substitution sites, and thermodynamic sequence-based features extracted from ProtDCal[28]. PON-tstab is a three-class predictor (stability increasing, decreasing, unchanged) and achieved the correct prediction ratio of around 0.5 versus the value 0.33 for a random predictor. This implies that, even with a data set of higher

quality, predicting protein stability remains an extremely challenging task[30]. Another intriguing application of ML in protein engineering is to design smart combinatorial libraries for directed protein evolution[**?**]. This has the potential to both reduce the experimental effort and improve the exploration of the sequence space by mutating multiple positions simultaneously. Moreover, it can approximate the empirical fitness landscape to suggest a refined set of variants for the next round of screening. Wu et al[**?**] used ML-assisted directed evolution to engineer an enzyme for a new stereodivergent carbonsilicon bond formation. The authors selected the reaction of phenyldimethyl silane with ethyl 2-diazopropanoate catalyzed by a putative nitric oxide dioxygenase from Rhodothermus marinus. They tested a variety of ML algorithms such as linear and kernel models, shallow neural networks, and ensemble methods to improve the enzyme enantioselectivity.

The linear regression and its variants were often used in the first attempts to obtain data-driven guidance, whereas lately there is a tendency to apply artificial neural networks and random forests, in part owing to the increase in data availability and improving high-throughput data collection methods

1.4.1. Current Challenges Related to ML-Aided

Methods. One of the main challenges in applications of ML to enzyme engineering stems from the intrinsic multi- disciplinarity of the approach. Biochemists, molecular biolo- gists, mathematicians, and computer scientists have to find a common language to clarify goals, carry out rigorous analysis and training, and avoid common pitfalls, wrongful usage of methods, and misinterpretations. Ready to use software packages certainly help standardize the training of an ML algorithm for nonspecialists, but heaping all the available data and running a range of ML algorithms to select the best predictor might not be the optimal strategy. The No Free Lunch theorem85 claims that no single ML method is superior to others a priori;86 therefore, a thorough understanding of the data types to be used and problems to be solved is essential in the development of efficient predictors. The current shift toward new and more complex ML methods, namely aggregating several algorithms into hybrid meta-predictors, hyperparameter optimization with many training subcycles, feature learning, and the fusion of ML-based and classical bioinformatics tools in a single predictor, will further challenge the crosstalk between disciplines necessary for the develop- ment of efficient and robust predictors in enzyme engineering. With the continuous growth of ML applications in enzyme engineering, the need for robust comparison of various predictors is of growing importance. This comparison is mainly obstructed by the lack of both standardized protocols for comparison and new data sets for testing. The lack of benchmark data sets, discrepancies in the performance measurements used, inaccurate or insufficient disclosure in publications, and the difficulty in finding reviewers with sufficiently broad expertise87 are among the most pressing issues. Researchers working on some applications with a long track record in bioinformatics, such as protein structure or function predictions, have already established several platforms that can be used for comparison of the ML predictors, i.e. CASP, CAFA, EFI, and COMBREX mentioned in section 4.1. Other applications have yet to see similar initiatives as, in our opinion, at least three key ingredients are necessary: (i) a sufficiently large community of researchers working on development of such applications, (ii) a sufficient amount of new high-quality data being collected regularly, and (iii) a leader that

will take on responsibility and invest time and effort into coordinating this activity. It is also worth noting that competitions of this kind are not flawless themselves, as their appearance led to an unwanted side effect: greater secrecy and an increased time delay before publishing newly developed methods due to the competition deadlines, which negatively affects the speed of knowledge circulation in science.

Few papers go beyond simple ROC analysis: e.g., resample cross-validation to estimate its statistical significance, explore the reasons for weak predictions, and analyze learning curves. Why does a particular predictor have a better performance? What features are critical for the performance of a predictor on a global scale? What ranges for feature values and what parts of the feature space are most critical for a particular data point to be classified correctly? Many articles on the topic lack this kind of analysis, which limits our understanding of the underlying molecular principles. In the next section, we touch upon modern trends in the ML workflow and architecture and also discuss how interpretable and explainable predictors can possibly provide some answers to the questions above. 4.3. Emerging Trends in ML-Based Methods for Enzyme Engineering. With the accumulation of more data by virtue of the emerging high-throughput experimental methods, the development of benchmark data sets and unified performance measurements is only a matter of time. Recently, an intriguing algorithm based on semisupervised learning has been presented to allow benchmarking in five different prediction tasks related to protein engineering, including secondary structure, fluorescence landscape, and stability landscape predictions.88 Moreover, as the data generation is streamlined, a data set from a single experiment is starting to have the size large enough for training ML algorithms to guide the design of future experiments, as was the case in the development of stereodivergent carbonsilicon bond forma- tion71 and the application of Gaussian processes to the directed evolution of cytochromes.89 The increase in the available data will prompt more extensive use of deep neural networks. This approach has already shown remarkable potential for complex tasks in genomics and proteomics but still has limited usage in enzyme engineering due to data scarcity. Sophisticated neural network architectures, such as recurrent or graph-based neural net- works, simultaneous training of several types of predictors (multitasking), combining structurally different input data (multimodal design), ML-based modeling of data sets (generative models), and retraining predictors used in one area by new data from another area (transfer learning) have only recently been applied in genomics.14 Several exciting attempts have also recently been made to apply some of those advanced techniques to proteins: using generative models to create soluble and functional malate dehydrogenase variants90 or predict mutational effects with high correlation with those actually observed in 42 high-throughput deep mutational scanning experiments.91 More data will also allow improving the existing methods, i.e. learning the optimal architecture of a predictor from the data (hyperparameter optimization),92 smart aggregation of several predictions from multiple methods,93 and introducing robust confidence scores for predictions.94 In enzyme engineering, this new level of algorithmic complexity will further save time and resources wasted on validating misleading predictions but will also require more sophisticated computer architecture, e.g. an increased use of parallel computing and stochastic training methods, which have already become standard techniques for the acceleration of deep neural network training.

We also envisage the combination of ML models with fundamentally different types of predictors. The development of hybrid methods became very successful, for example, in the

prediction of protein stability.5 Moreover, models targeting several proper- ties of biocatalyst simultaneously, e.g. activity, stability, and solubility, would dramatically reduce the risk of unsuccessful laboratory experiments resulting from in silico design of active but unstable or poorly soluble proteins. Another noticeable trend in ML is toward interpretable and explainable predictors.95 Apart from the global importance of features for ML predictors, feature importance scores calculated for each input example96,97 may help explain why a particular prediction was made for each input data point. In addition to providing mechanistic insights, interpretable algorithms can aid in smart biocatalyst design. For instance, instead of simply screening all the possible mutations with an ML-based tool to improve a target property, researchers can make use of designing variants on the basis of the structure of a predictor using adaptive sampling.98 Such an approach favors predictors whose parameters can provide such guidance: e.g., linear predictors over more flexible yet harder to interpret artificial neural networks (Figure 3). Linear predictors allow analytical design on the basis of the coefficients;99 in contrast, sophisticated predictors are usually prone to pathological behavior, i.e. sudden misclassification after a slight and almost imperceptible perturbation of input.100 Another promising approach is to use interpretable architectures of predictors already at the design stage, e.g. the visible neural networks.101 The design of such networks is guided by the knowledge of the underlying biological mechanism, e.g. the choice of layers and the connections between layers may mimic the hierarchical organization of transcriptional regulatory factors in the cell nucleus.

2. Artículo 2:

3. Referencias

Referencias

[1] Mazurenko, S., Prokop, Z., and Damborsky, J. (2019). Machine learning in enzyme engineering. ACS Catalysis, 10(2), 1210-1223.

[2] Yang, Y.; Niroula, A.; Shen, B.; Vihinen, M. PON-Sol: Prediction of Effects of Amino Acid Substitutions on Protein Solubility. Bioinformatics 2016, 32, 2032!2034.

[3] Folkman, L.; Stantic, B.; Sattar, A.; Zhou, Y. EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. J. Mol. Biol. 2016, 428, 1394! 1405.

[4] Teng, S.; Srivastava, A. K.; Wang, L. Sequence Feature-Based Prediction of Protein Stability Changes upon Amino Acid Substitutions. BMC Genomics 2010, 11, S5.

[5] Huang, L.; Gromiha, M. M.; Ho, S. iPTREE-STAB: Interpretable Decision Tree Based Method for Predicting Protein Stability Changes Upon Mutations. Bioinformatics 2007, 23, 1292! 1293.

[6] Koskinen, P.; Toronen, P.; Nokso-Koivisto, J.; Holm, L. PANNZER: High-Throughput Functional Annotation of Uncharac- terized Proteins in an Error-Prone Environment. Bioinformatics 2015, 31, 1544!1552.

[7] De Ferrari, L.; Mitchell, J. B. From Sequence to Enzyme Mechanism Using Multi-Label Machine Learning. BMC Bioinf. 2014, 15, 150.

[8] Falda, M.; Toppo, S.; Pescarolo, A.; Lavezzo, E.; Di Camillo, B.; Facchinetti, A.; Cilia, E.; Velasco, R.; Fontana, P. Argot2: A Large Scale Function Prediction Tool Relying on Semantic Similarity of Weighted Gene Ontology Terms. BMC Bioinf. 2012, 13, S14.

[9] Cozzetto, D.; Buchan, D. W.; Bryson, K.; Jones, D. T. Protein Function Prediction by Massive Integration of Evolutionary Analyses and Multiple Data Sources. BMC Bioinf. 2013, 14, S1.

[10] Kulski, J. Next Generation Sequencing: Advances, Applications and Challenges; InTechOpen: London, 2016.

[11] Straiton, J.; Free, T.; Sawyer, A.; Martin, J. From Sanger Sequencing to Genome Databases and Beyond. BioTechniques 2019, 66, 6063.

[12] Ardui, S.; Ameur, A.; Vermeesch, J. R.; Hestand, M. S. Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics. Nucleic Acids Res. 2018, 46, 21592168.

[13] Kono, N.; Arakawa, K. Nanopore Sequencing: Review of Potential Applications in Functional Genomics. Dev., Growth Differ. 2019, 61, 316326

[14] Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D. Speeding Up Enzyme Discovery and Engineering with Ultrahigh-Throughput Methods. Curr. Opin. Struct. Biol. 2018, 48, 149156.

[15] Wrenbeck, E. E.; Faber, M. S.; Whitehead, T. A. Deep Sequencing Methods for Protein Engineering and Design. Curr. Opin. Struct. Biol. 2017, 45, 3644.

[16] Fowler, D. M.; Fields, S. Deep Mutational Scanning: A New Style of Protein Science. Nat. Methods 2014, 11, 801807.

[17] Gupta, K.; Varadarajan, R. Insights into Protein Structure, Stability and Function from Saturation Mutagenesis. Curr. Opin. Struct. Biol. 2018, 50, 117125.

[18] UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. Nucleic Acids Res. 2018, 47, D506D515.

[19] Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Zidek, A.; Nelson, A.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D.; Senior, A. W. De Novo Structure Prediction with Deeplearning Based Scoring. In Thirteenth Critical Assessment of Techniques for Protein Structure Prediction Abstracts; 2018; pp 1112.

[20] Kinch, L. N.; Shi, S.; Cheng, H.; Cong, Q.; Pei, J.; Mariani, V.; Schwede, T.; Grishin, N. V. CASP9 Target Classification. Proteins: Struct., Funct., Genet. 2011, 79, 2136.

[21] Shehu, A.; Barbará, D.; Molloy, K. A Survey of ComputationalMethods for Protein Function Prediction. In Big Data Analytics in Genomics; Wong, K. C., Ed.; Springer: Cham, 2016; pp 225298.

[22] Zhang, C.; Freddolino, P. L.; Zhang, Y. COFACTOR: Improved Protein Function Prediction by Combining Structure,Sequence and ProteinProtein Interaction Information. Nucleic Acids Res. 2017, 45, W291W299.

[23] Kumar, N.; Skolnick, J. EFICAz2. 5: Application of a High-Precision Enzyme Function Predictor to 396 Proteomes. Bioinformatics 2012, 28, 26872688.

[24] Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. Bioinformatics 2018, 34, 760769.

[25] Yang, M.; Fehl, C.; Lees, K. V.; Lim, E. K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G. Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. Nat. Chem. Biol. 2018, 14, 11091117.

[26] Niwa, T.; Ying, B. W.; Saito, K.; Jin, W.; Takada, S.; Ueda, T.; Taguchi, H. Bimodal Protein Solubility Distribution Revealed by an Aggregation Analysis of the Entire Ensemble of Escherichia Coli Proteins. Proc. Natl. Acad. Sci. U. S. A. 2009, 106, 42014206.

[27] Klesmith, J. R.; Bacik, J. P.; Wrenbeck, E. E.; Michalczyk, R.; Whitehead, T. A. Trade-Offs Between Enzyme Fitness and Solubility Illuminated by Deep Mutational Scanning. Proc. Natl. Acad. Sci. U. S. A. 2017, 114, 22652270.

[28] Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y. ProtDCal: A Program to Compute General-Purpose-Numerical Descriptors for Sequences and 3D-Structures of Proteins. BMC Bioinf. 2015, 16, 162.

[29] Han, X.; Wang, X.; Zhou, K. Develop Machine Learning-Based Regression Predictive Models for Engineering Protein Solubility. Bioinformatics 2019, 35, 46404646.

[30] Musil, M.; Konegger, H.; Hon, J.; Bednar, D.; Damborsky, J. Computational Design of Stable and Soluble Biocatalysts. ACS Catal. 2019, 9, 10331054.