

RUDEUS: A Machine Learning Classification System to Study DNA-Binding Proteins

David Medina-Ortiz^{1,2}, Gabriel Cabas-Mora¹, Iván Moya¹, Nicole Soto-García¹
and Roberto Uribe-Paredes¹

¹Departamento de Ingeniería En Computación, Universidad de Magallanes, Avenida Bulnes 01855, Punta Arenas, Chile

²Centre for Biotechnology and Bioengineering, CeBiB, Universidad de Chile, Beauchef 851, Santiago, Chile

Keywords: DNA-Binding Proteins, Single-Stranded and Double-Stranded DNA, Machine Learning, Protein Language Models.

Abstract: DNA-binding proteins play crucial roles in biological processes such as replication, transcription, packaging, and chromatin remodeling. Their study has gained importance across scientific fields, with computational biology complementing traditional methods. While machine learning has advanced bioinformatics, generalizable pipelines for identifying DNA-binding proteins and their specific interactions remain scarce. We present RUDEUS, a Python library with hierarchical classification models to identify DNA-binding proteins and distinguish between single- and double-stranded DNA interactions. RUDEUS integrates protein language models, supervised learning, and Bayesian optimization, achieving 95% precision in DNA-binding identification and 89% accuracy in distinguishing interaction types. The library also includes tools for annotating unknown sequences and validating DNA-protein interactions through molecular docking. RUDEUS delivers competitive performance and is easily integrated into protein engineering workflows. It is available under the MIT License, with the source code and models available on the GitHub repository <https://github.com/ProteinEngineering-PESB2/RUDEUS>.

1 INTRODUCTION


DNA-protein interactions are fundamental to numerous cellular processes critical for biological functions. Approximately 6-7% of eukaryotic proteins interact with DNA, utilizing specific DNA-binding domains and varying affinities for single- and double-stranded DNA (Attali et al., 2021; Gupta et al., 2021). These interactions are driven by direct base–amino acid recognition and indirect forces from DNA conformational changes (Arora et al., 2023).


DNA-binding proteins (DBPs) play key roles in processes like DNA replication, transcription, packaging, and chromatin remodeling (Kabir et al., 2024). They aid in strand separation, maintain DNA integrity, regulate gene expression, and influence chromatin structure. Understanding DBPs is essential for insights into gene regulation and links between


mutations and genetic diseases (Zhang et al., 2022; Kabir et al., 2024). Recent studies on proteins such as TDP-43 and helicase chromodomain proteins have advanced knowledge in fields like neurodegeneration and cancer (Lye and Chen, 2022; Alendar and Berns, 2021; Wang et al., 2022a).


Computational biology, bolstered by AI and machine learning, has enhanced the discovery of DBPs by predicting interaction sites and transcription factor binding hotspots (Wang et al., 2022b). While many machine learning models have been applied, including deep learning, comparing them is difficult due to variations in datasets and validation methods (Shadab et al., 2020; Zhang et al., 2020; Ali et al., 2022; Banjar et al., 2022; Barukab et al., 2022). Recent approaches have employed large protein language models for more robust numerical representations (Medina et al., 2023; Medina-Ortiz et al., 2024; Fernández et al., 2023).

This paper introduces RUDEUS, a Python library designed for DNA-binding classification and distinguishing between single- and double-stranded interactions. RUDEUS combines protein language mod-

^a <https://orcid.org/0000-0002-8369-5746>

^b <https://orcid.org/0009-0004-2344-9860>

^c <https://orcid.org/0000-0002-0458-378X>

^d <https://orcid.org/0009-0001-1438-1938>

els, supervised learning algorithms, and Bayesian hyperparameter tuning to build predictive models. Achieving precision rates of 95% for DNA-binding identification and 89% for interaction type evaluation, RUDEUS demonstrates strong performance. It annotated over 20,000 protein sequences and was validated using structural bioinformatics. The library's flexibility and ease of use make it a valuable tool for exploring latent space and mutation landscapes in DBPs.

2 METHODS

2.1 Collecting and Processing Protein Sequences

All protein sequences were sourced from the literature, including datasets from Hu et al. (2019); Shadab et al. (2020); Sharma et al. (2021); Wang et al. (2017). After collection, a preprocessing step was applied to merge, clean, and remove redundancy and inconsistencies. Filters were then applied to exclude non-canonical sequences and select sequences within a length range of 50 to 1024 amino acids. Additionally, homology redundancy was eliminated using the CDHit library Fu et al. (2012).

2.2 Numerical Representation Strategies

This work explore different pre-trained models based on protein language models, including ProTrans (El-naggar et al., 2020) and ESM (Rives et al., 2021; Meier et al., 2021). All pre-trained models were applied through the bio-embedding tool, combined with a reduction process to obtain vectors in a $1 - D$ dimension (Dallago et al., 2021). Moreover, physicochemical based approaches and Fourier transforms also were explored (Medina-Ortiz et al., 2022, 2020a).

2.3 Training Predictive Models and Tuning Optimization

A classic machine learning pipeline was employed to train predictive models Medina-Ortiz et al. (2020b). The datasets were first split into training (70%), validation (20%), and testing (10%) sets. The models were then trained using the strategies proposed in Medina-Ortiz et al. (2024), which included an exploration phase, statistical methods to select the best combinations of numerical representation strategies and machine learning algorithms, and Bayesian approaches for hyperparameter tuning Akiba et al.

(2019). Once the models were trained, the testing datasets were used for benchmarking, and the models were deployed to predict unknown protein sequences (See Figure 3 of Appendix for a schematic representation of the employed pipeline to train the predictive models).

2.4 Structural Bioinformatics Approaches

RUDEUS incorporates a structural bioinformatics pipeline to validate model predictions using DNA-protein molecular docking via LightDock v9.4 Roel-Touris et al. (2020). The pipeline prepares protein structures by applying protonation, hydrogen deletion, structure rebuilding with the Reduce library, and modifying atoms to comply with the AMBER94 force field. After preparation, molecular docking is performed with 400 swarms, 200 glowworms, and 100 steps. The resulting conformers are clustered using the RMSD metric with the BSAS function, and the best pose is selected based on the highest docking score.

2.5 Availability and Implementation Strategies

All source code was implemented under the Python Language programming v3.9.16, including the modules, libraries, and demonstration scripts in RUDEUS. The main libraries employed to develop the predictive models were scikit-learn (Pedregosa et al., 2011) and Optuna (Akiba et al., 2019). Furthermore, to process and compile all datasets, the Pandas library was employed (McKinney et al., 2011). Finally, a conda environment was constructed to facilitate the deployment of the built library, combined with different Jupyter Notebooks, to ensure the replicability of the presented work. All source code, environment configuration, datasets, and created models are available for non-commercial uses in the GitHub repository under the MIT licence <https://github.com/ProteinEngineering-PESB2/RUDEUS>.

3 RESULTS AND DISCUSSIONS

3.1 RUDEUS Achieves High Performances in Its Classification Models

Two classification tasks were explored in RUDEUS: DNA-binding protein classification and the identifi-

cation of single- versus double-stranded DNA interactions. For each task, over 10,000 combinations of numerical representation strategies and supervised learning algorithms were evaluated. The models' performance was measured using accuracy, precision, recall, and F-score.

Figure 4 of Appendix displays the recall metric distributions for the training process. On average, the models achieved 83% precision for DNA-binding classification and 82% precision for DNA strand type prediction. The highest-performing DNA-binding models were based on pre-trained ProtTrans Uniref, BDF, and XLU50 models, independent of the learning algorithm. For DNA strand type classification, the best results came from ProtTrans XLU50, Uniref, t5bdf, ESM1B, and ESM1V models. Ensemble methods like Random Forest, Gradient Boosting, ExtraTrees, and KNN consistently delivered the top results for both tasks.

A statistical selection process identified the best combinations of representation strategies and algorithms. Sixteen Bernoulli events were evaluated using two filters: i) top-performing models above the 90th quantile and ii) models with standard deviations below the 10th quantile. A binomial distribution was then applied to detect outliers, with a success threshold of > 12 events, representing a success probability below 0.01. This stringent selection yielded five optimal combinations for DNA-binding classification and four for DNA strand type prediction, as summarized in Table 1. While the selected models exhibited strong performance, overfitting was observed, with differences between training and validation metrics.

Table 1: Selected combinations of supervised learning algorithms and numerical representation approaches for all tasks explored in this work.

Task	Algorithm	Encoder	Recall
DNA-binding classification	ExtraTrees	prot. Uniref	0.93
	ExtraTrees	prot. bdf	0.93
	Gradient B	prot. Uniref	0.91
	KNeighbors	prot. Uniref	0.93
	RandomForest	prot. Uniref	0.93
	ExtraTrees	prot. Uniref	0.90
Single-stranded or double-stranded	ExtraTrees	prot. XLU50	0.90
	Gaussian Pro.	prot. XLU50	0.89
	SVC	prot. XLU50	0.90
	ExtraTrees	prot. XLU50	0.90

All selected combinations of supervised learning algorithms and numerical representation strategies were optimized using the Optuna library (Akiba et al., 2019). Two models were then selected based on the criteria outlined in the pipeline. For DNA-binding prediction, the ExtraTrees algorithm combined with the ProtTrans Uniref model was chosen, while for single-stranded or double-stranded DNA in-

teraction, the same algorithm was used, but the ProtTrans XLU50 model was selected. The DNA-binding model achieved 95% precision with a Matthews correlation coefficient (MCC) of 0.89, and the single-stranded/double-stranded model achieved 89% precision with an MCC of 0.81.

Figure 1 summarizes both models' performance. The confusion matrices (Figure 1 A and 1 C) indicate strong performance in identifying positive and negative classes, with the DNA-binding model outperforming the single/double-stranded model in distinguishing interactions. Precision-recall curves (Figure 1 B and 1 D) showed average precision values of 0.98 and 0.96, respectively, aligning with the confusion matrices and demonstrating the greater difficulty in classifying interaction types. ROC curves, calculated using $k = 5$ cross-validation, revealed area under the curve (AUC) scores of 0.98 for DNA-binding and 0.97 for the interaction model, confirming the models' robust predictive capabilities.

Table 2 compares the RUDEUS models with state-of-the-art methods. For DNA-binding, RUDEUS achieved the highest specificity (95.5%) and MCC (0.89), while the method in (Zhang et al., 2021) had the highest sensitivity, differing only by 0.1% from RUDEUS. For the single-stranded/double-stranded task, RUDEUS achieved the highest MCC (0.81), although other methods reported higher sensitivity (Ali et al., 2020) and specificity (Tan et al., 2019). However, these methods showed signs of overfitting, as indicated by large gaps between sensitivity and specificity and lower MCC values compared to RUDEUS.

Table 2: State-of-the-art comparison for DNA-binding classification models and single-stranded or double-stranded interaction models.

Task	Classifier	SN(%)	SP(%)	MCC	Reference
DNA-binding	RF	79.3	89.0	0.69	(Kumar et al., 2009)
	RF	83.7	90.0	0.72	(Ma et al., 2016)
	SVM	87.0	85.5	0.72	(Zaman et al., 2017)
	SVM	89.1	88.8	0.78	(Ali et al., 2018)
	SVM	94.1	97.6	0.92	(Rahman et al., 2018)
	SVM	91.1	88.8	0.79	(Mishra et al., 2019)
	SVM	91.8	93.0	0.84	(Ali et al., 2019)
	SVM	93.4	93.4	0.86	(Zhang et al., 2021)
	ExtraTrees	93.3	95.5	0.89	This work
	RF	90.8	78.8	0.64	(Wang et al., 2017)
Single-stranded or double-stranded	SVM	94.2	80.33	0.72	(Ali et al., 2020)
	GTB	78.4	97.5	0.79	(Tan et al., 2019)
	HMM	85.3	92.8	0.78	(Sharma et al., 2021)
	ExtraTrees	87.8	91.6	0.81	This work

3.2 RUDEUS Facilitate the Exploration of Single-Stranded or Double-Stranded Interaction Evaluation

More than 20,000 DNA-binding protein sequences were classified as either single- or double-stranded

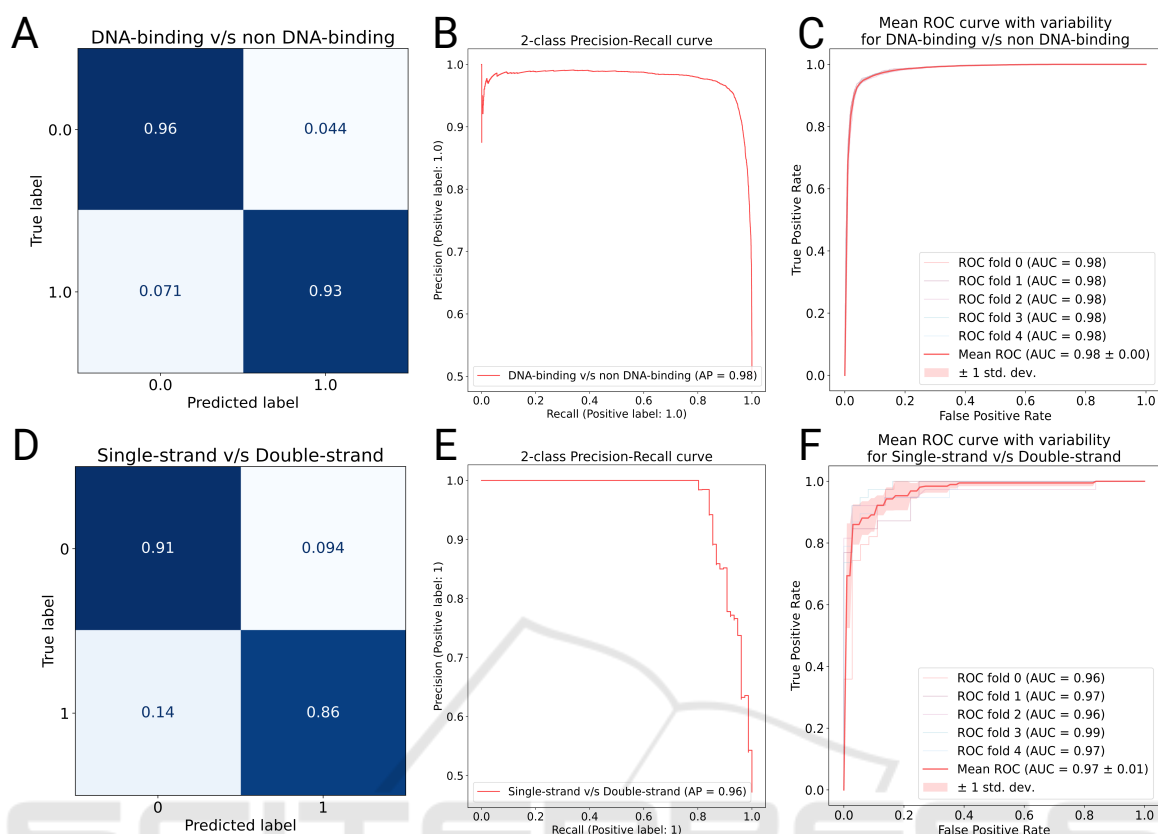


Figure 1: Description through different performances visualization the selected and optimized models for both tasks explored in this work. A-D Confusion matrix estimated during the validation process for DNA-binding task single-stranded or double-stranded task, respectively. **B-E** Precision-recall curve estimated during the validation process for DNA-binding task single-stranded or double-stranded task, respectively. The average precision (AP) was calculated in both cases, achieving 0.98 and 0.96, respectively. **C-F** Receiver operating characteristic (ROC) curve estimated during the training process for DNA-binding task single-stranded or double-stranded task, respectively. In both cases, the area under the curve (AUC) was estimated to achieve 0.98 and 0.97, respectively.

using the exploration module in RUDEUS. First, the sequences were numerically represented using pre-trained models selected for strand interaction classification. The predictions showed that over 18,000 proteins were classified as double-stranded, while around 2,000 were identified as single-stranded, reflecting proportions similar to the dataset used for model training.

Three DNA-binding proteins with identified strand interactions were further evaluated using the bioinformatics structural pipeline. Figure 2 provides molecular docking visualizations and detailed interaction site analyses for these proteins, all of which were previously reported in the literature.

Figure 2 A illustrates the molecular docking of protein 1BNZ, a hyperthermophile chromosomal protein that binds double-stranded DNA (Gao et al., 1998; Guagliardi et al., 2002). Key hydrophobic residues—TRP24, VAL26, MET29, and

ALA45—play a significant role in DNA binding (Figure 2 B). Interactions occur via hydrogen bonds, salt bridges, and van der Waals contacts, consistent with previous reports (Gao et al., 1998).

Similarly, Figure 2 C shows the docking of protein 1HRY, which is involved in sexual differentiation by regulating the gene responsible for Müllerian duct regression in male embryos (Werner et al., 1995). Six residues (ASN10, PHE12, ILE13, SER33, ILE35, SER36, TYR74) interact with DNA bases, forming hydrogen bonds and electrostatic interactions (Figure 2 D), as described in (Werner et al., 1995).

In contrast, Figure 2 E presents the docking of protein 3ULP, known as Pf-SSB, a single-stranded DNA-binding protein crucial for DNA metabolism in the malaria-causing parasite (Antony et al., 2012). The homotetramer structure of 3ULP features identical DNA-contacting residues (S110, N114, T129) across all four subunits (Figure 2 F), which form part

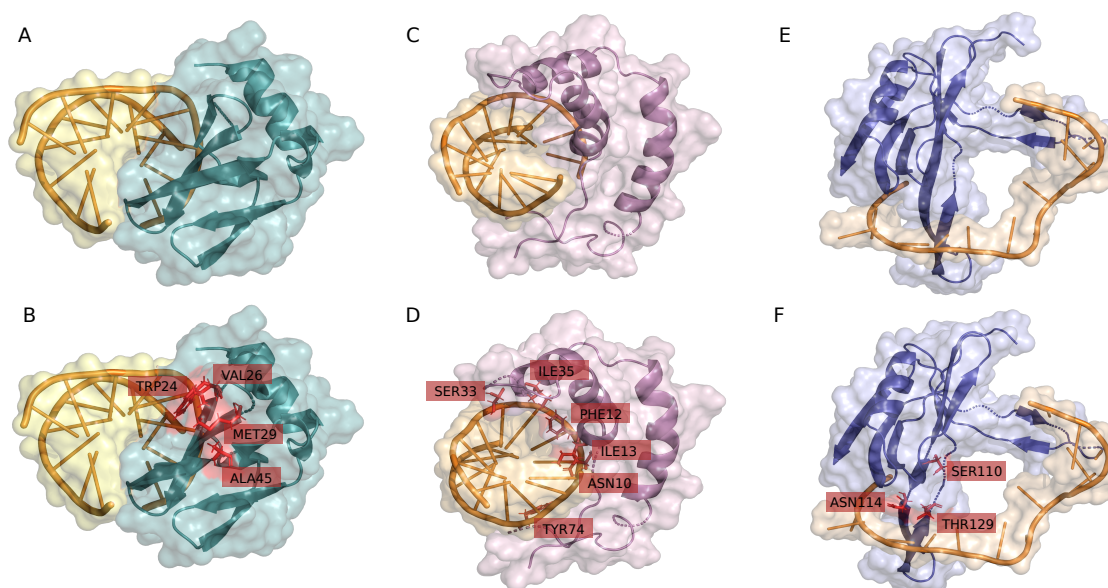


Figure 2: **Structural bioinformatics validation through DNA-protein molecular docking for three DNA-binding proteins and their interaction type identified with the models available in RUDEUS.** A-B DNA-protein molecular docking and the most relevant identified residues for the DNA interaction for the protein 1BNZ. C-D DNA-protein molecular docking and the most relevant identified residues for the interaction for the protein 1HRY. E-F DNA-protein molecular docking and the most relevant identified residues for the interaction for the protein 3ULP.

of the replication and maintenance machinery in the apicoplast (Antony et al., 2012).

4 CONCLUSIONS

This work introduces RUDEUS, a Python library specifically designed for the investigation and classification of DNA-binding proteins, as well as the identification of DNA strand interaction types. The methodology incorporates a flexible pipeline that leverages protein language models, supervised learning algorithms, and Bayesian optimization to train high-performance classification models. These models surpass state-of-the-art benchmarks in sensitivity, specificity, and MCC scores, demonstrating RUDEUS's superiority in this domain, while maintaining the simplicity and replicability of existing methods.

An extensive exploration process highlighted the utility of RUDEUS, enabling the annotation of over 20,000 protein sequences as single- or double-stranded, validated through structural bioinformatic approaches and DNA-protein molecular docking. RUDEUS's intuitive interface and powerful features make it highly applicable for integration into broader protein design pipelines, including landscape reconstruction, directed evolution, and latent space explo-

ration using deep generative models.

COMPETING INTERESTS

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS STATEMENT

IM-B and DM-O: conceptualization. DM-O, GC-M, and NS-G: methodology. DM-O and RU-P: validation. IM-B, GC-M, and NS-G: investigation. DM-O, IM-B, RU-P, and GC-M: writing, review, and editing. DM-O and RU-P: supervision and funding resources. DM-O: project administration.

ACKNOWLEDGEMENTS

This research has been financed mainly by the Centre for Biotechnology and Bioengineering - CeBiB (PIA project FB0001, Conicyt, Chile). DM-O acknowledges ANID for the project "SUBVENCIÓN A IN-

STALACIÓN EN LA ACADEMIA CONVOCATORIA AÑO 2022”, Folio 85220004. RU-P acknowledges ANID for the grant Fondecyt 1230298.

REFERENCES

- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.
- Alendar, A. and Berns, A. (2021). Sentinels of chromatin: chromodomain helicase dna-binding proteins in development and disease. *Genes & Development*, 35(21-22):1403–1430.
- Ali, F., Ahmed, S., Swati, Z. N. K., and Akbar, S. (2019). Dp-binder: machine learning model for prediction of dna-binding proteins by fusing evolutionary and physicochemical information. *Journal of Computer-Aided Molecular Design*, 33:645–658.
- Ali, F., Arif, M., Khan, Z. U., Kabir, M., Ahmed, S., and Yu, D.-J. (2020). Sdbp-pred: Prediction of single-stranded and double-stranded dna-binding proteins by extending consensus sequence and k-segmentation strategies into pssm. *Analytical biochemistry*, 589:113494.
- Ali, F., Kabir, M., Arif, M., Swati, Z. N. K., Khan, Z. U., Ullah, M., and Yu, D.-J. (2018). Dbppred-pdsd: Machine learning approach for prediction of dna-binding proteins using discrete wavelet transform and optimized integrated features space. *Chemometrics and Intelligent Laboratory Systems*, 182:21–30.
- Ali, F., Kumar, H., Patil, S., Ahmed, A., Banjar, A., and Daud, A. (2022). Dbp-deepcnn: prediction of dna-binding proteins using wavelet-based denoising and deep learning. *Chemometrics and Intelligent Laboratory Systems*, 229:104639.
- Antony, E., Weiland, E. A., Korolev, S., and Lohman, T. M. (2012). Plasmodium falciparum ssb tetramer wraps single-stranded dna with similar topology but opposite polarity to e. coli ssb. *Journal of molecular biology*, 420(4-5):269–283.
- Arora, S., Gupta, S., Verma, S., and Malik, I. (2023). Prediction of dna interacting residues. In *2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, pages 54–57. IEEE.
- Attali, I., Botchan, M. R., and Berger, J. M. (2021). Structural mechanisms for replicating dna in eukaryotes. *Annual review of biochemistry*, 90:77–106.
- Banjar, A., Ali, F., Alghushairy, O., and Daud, A. (2022). idbp-pbmd: A machine learning model for detection of dna-binding proteins by extending compression techniques into evolutionary profile. *Chemometrics and Intelligent Laboratory Systems*, 231:104697.
- Barukab, O., Ali, F., Alghamdi, W., Bassam, Y., and Khan, S. A. (2022). Dbp-cnn: Deep learning-based prediction of dna-binding proteins by coupling discrete cosine transform with two-dimensional convolutional neural network. *Expert Systems with Applications*, 197:116729.
- Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., and Rost, B. (2021). Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113.
- Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2020). Prottrans: Towards cracking the language of life’s code through self-supervised deep learning and high performance computing.
- Fernández, D., Olivera-Nappa, Á., Uribe-Paredes, R., and Medina-Ortiz, D. (2023). Exploring machine learning algorithms and protein language models strategies to develop enzyme classification systems. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 307–319. Springer.
- Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.
- Gao, Y.-G., Su, S.-Y., Robinson, H., Padmanabhan, S., Lim, L., McCrary, B. S., Edmondson, S. P., Shriver, J. W., and Wang, A. H.-J. (1998). The crystal structure of the hyperthermophile chromosomal protein sso7d bound to dna. *Nature structural biology*, 5(9):782–786.
- Guagliardi, A., Cerchia, L., Rossi, M., et al. (2002). The sso7d protein of *Sulfolobus solfataricus*: in vitro relationship among different activities. *Archaea*, 1:87–93.
- Gupta, N. K., Wilkinson, E. A., Karuppanan, S. K., Bailey, L., Vilan, A., Zhang, Z., Qi, D.-C., Tadich, A., Tuite, E. M., Pike, A. R., et al. (2021). Role of order in the mechanism of charge transport across single-stranded and double-stranded dna monolayers in tunnel junctions. *Journal of the American Chemical Society*, 143(48):20309–20319.
- Hu, S., Ma, R., and Wang, H. (2019). An improved deep learning method for predicting dna-binding proteins based on contextual features in amino acid sequences. *PLoS one*, 14(11):e0225317.
- Kabir, A., Bhattarai, M., Rasmussen, K. O., Shehu, A., Bishop, A. R., Alexandrov, B. S., and Usheva, A. (2024). Advancing transcription factor binding site prediction using dna breathing dynamics and sequence transformers via cross attention. *bioRxiv*, pages 2024–01.
- Kumar, K. K., Pugalenthi, G., and Suganthan, P. N. (2009). Dna-prot: identification of dna binding proteins from protein sequence information using random forest. *Journal of Biomolecular Structure and Dynamics*, 26(6):679–686.
- Lye, Y. S. and Chen, Y.-R. (2022). Tar dna-binding protein 43 oligomers in physiology and pathology. *IUBMB life*, 74(8):794–811.
- Ma, X., Guo, J., and Sun, X. (2016). Dnabp: Identification of dna-binding proteins based on feature selection using a random forest and predicting binding residues. *PLoS one*, 11(12):e0167345.

- McKinney, W. et al. (2011). pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.
- Medina, D., Sepulveda-Yanez, J., Alvarez-Saravia, D., Uribe-Paredes, R., Veelken, H., and Navarrete, M. (2023). Artificial intelligence approach for the discovery of autoantigen recognition by b-cell lymphomas. *Blood*, 142:125.
- Medina-Ortiz, D., Contreras, S., Amado-Hinojosa, J., Torres-Almonacid, J., Asenjo, J. A., Navarrete, M., and Olivera-Nappa, A. (2020a). Combination of digital signal processing and assembled predictive models facilitates the rational design of proteins. *arXiv preprint arXiv:2010.03516*.
- Medina-Ortiz, D., Contreras, S., Amado-Hinojosa, J., Torres-Almonacid, J., Asenjo, J. A., Navarrete, M., and Olivera-Nappa, A. (2022). Generalized property-based encoders and digital signal processing facilitate predictive tasks in protein engineering. *Frontiers in Molecular Biosciences*, 9.
- Medina-Ortiz, D., Contreras, S., Fernández, D., Soto-García, N., Moya, I., Cabas-Mora, G., and Olivera-Nappa, A. (2024). Protein language models and machine learning facilitate the identification of antimicrobial peptides. *International Journal of Molecular Sciences*, 25(16):8851.
- Medina-Ortiz, D., Contreras, S., Quiroz, C., and Olivera-Nappa, A. (2020b). Development of supervised learning predictive models for highly non-linear biological, biomedical, and general datasets. *Frontiers in molecular biosciences*, 7:13.
- Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc.
- Mishra, A., Pokhrel, P., and Hoque, M. T. (2019). Stackdpred: a stacking based prediction of dna-binding protein from sequence. *Bioinformatics*, 35(3):433–441.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.
- Rahman, M. S., Shatabda, S., Saha, S., Kaykobad, M., and Rahman, M. S. (2018). Dpp-pseaac: a dna-binding protein prediction model using chou’s general pseaac. *Journal of theoretical biology*, 452:22–34.
- Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15).
- Roel-Touris, J., Bonvin, A. M., and Jiménez-García, B. (2020). Lightdock goes information-driven. *Bioinformatics*, 36(3):950–952.
- Shadab, S., Khan, M. T. A., Neezi, N. A., Adilina, S., and Shatabda, S. (2020). Deepdbp: deep neural networks for identification of dna-binding proteins. *Informatics in Medicine Unlocked*, 19:100318.
- Sharma, R., Kumar, S., Tsunoda, T., Kumarevel, T., and Sharma, A. (2021). Single-stranded and double-stranded dna-binding protein prediction using hmm profiles. *Analytical biochemistry*, 612:113954.
- Tan, C., Wang, T., Yang, W., and Deng, L. (2019). Predpsd: a gradient tree boosting approach for single-stranded and double-stranded dna binding protein prediction. *Molecules*, 25(1):98.
- Wang, W., Sun, L., Zhang, S., Zhang, H., Shi, J., Xu, T., and Li, K. (2017). Analysis and prediction of single-stranded and double-stranded dna binding proteins based on protein sequences. *BMC bioinformatics*, 18:1–10.
- Wang, Y., Zhang, L., Huang, T., Wu, G.-R., Zhou, Q., Wang, F.-X., Chen, L.-M., Sun, F., Lv, Y., Xiong, F., et al. (2022a). The methyl-cpg-binding domain 2 facilitates pulmonary fibrosis by orchestrating fibroblast to myofibroblast differentiation. *European Respiratory Journal*, 60(3).
- Wang, Z., Gong, M., Liu, Y., Xiong, S., Wang, M., Zhou, J., and Zhang, Y. (2022b). Towards a better understanding of tf-dna binding prediction from genomic features. *Computers in Biology and Medicine*, 149:105993.
- Werner, M. H., Huth, J. R., Gronenborn, A. M., and Clore, G. M. (1995). Molecular basis of human 46x, y sex reversal revealed from the three-dimensional solution structure of the human sry-dna complex. *Cell*, 81(5):705–714.
- Zaman, R., Chowdhury, S. Y., Rashid, M. A., Sharma, A., Dehzangi, A., Shatabda, S., et al. (2017). Hmmbinder: Dna-binding protein prediction using hmm profile based features. *BioMed research international*, 2017.
- Zhang, J., Chen, Q., and Liu, B. (2020). idrbp_mmc: identifying dna-binding proteins and rna-binding proteins based on multi-label learning model and motif-based convolutional neural network. *Journal of molecular biology*, 432(22):5860–5875.
- Zhang, Q., Liu, P., Wang, X., Zhang, Y., Han, Y., and Yu, B. (2021). Stackpdb: predicting dna-binding proteins based on xgb-rfe feature optimization and stacked ensemble classifier. *Applied Soft Computing*, 99:106921.
- Zhang, Y., Bao, W., Cao, Y., Cong, H., Chen, B., and Chen, Y. (2022). A survey on protein–dna-binding sites in computational biology. *Briefings in Functional Genomics*, 21(5):357–375.

APPENDIX

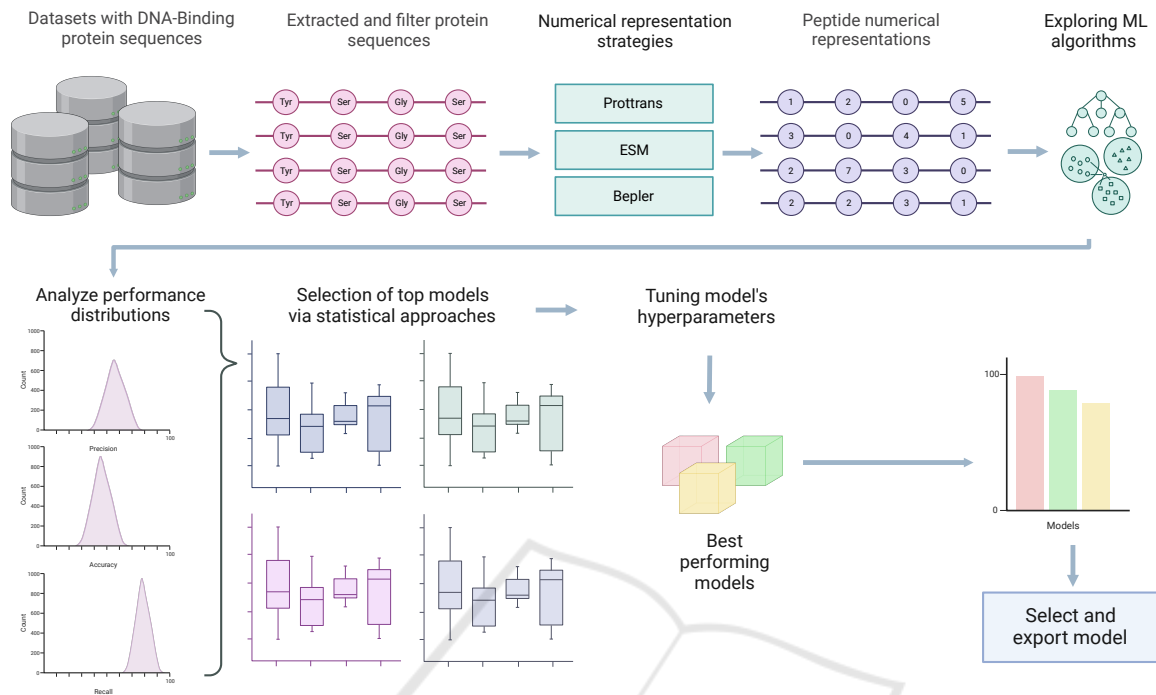


Figure 3: **The designed e implemented pipeline to train predictive models for DNA-Binding identification incorporated in RUDEUS.** The proposed pipeline first collects and processes the protein sequences by incorporating length filters and removing non-canonical residues. Then, numerical representation strategies are applied to obtain encoded vectors through pre-trained models based on protein language models, including Prottrans family models, ESM family models, Bepler, Glove, and all the different pre-trained models available in the bio-embedding library. Then, different supervised learning algorithms are explored using default hyperparameters employing all generated datasets in the previous step. Then, statistical approaches are applied to filter and select the best combinations of supervised learning algorithms and numerical representation approaches. A Bayesian approach guides the selected combinations tuning hyperparameters process through the Optuna library, and ensemble learning is explored to evaluate different combinations of the individual optimized models. Finally, the best strategy is selected based on the best performances, including training, validation, and overfitting ratio.

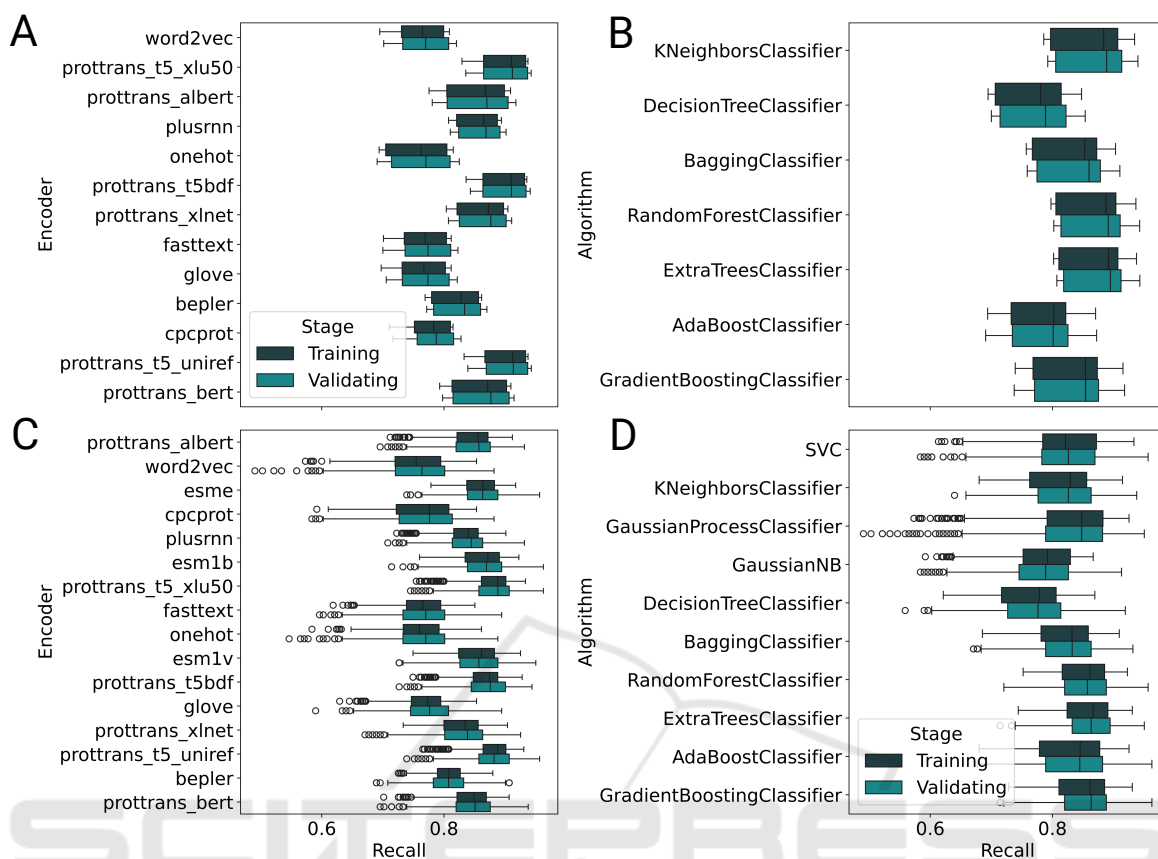


Figure 4: **Recall distribution performances for all explored tasks in this work evaluated by numerical representation strategies and supervised learning algorithms.** **A** Recall distribution for DNA-binding classification task grouped by pre-trained model employed as numerical representation strategy. **B** Recall distribution for DNA-binding classification task grouped by supervised learning algorithm. **C** Recall distribution for single-stranded or double-stranded DNA type interaction task grouped by pre-trained model employed as numerical representation strategy. **D** Recall distribution for single-stranded or double-stranded DNA type interaction task grouped by supervised learning algorithm.