Variational Autoencoders and Evolutionary Algorithms for Targeted Novel Enzyme Design

Miguel Martins
Department of Informatics
University of Minho, Portugal
mmiguelawes@gmail.com

Miguel Rocha
Centre of Biological Engineering
Department of Informatics
University of Minho, Portugal
LABBELS – Associate Laboratory,
Braga/Guimarães, Portugal
mrocha@di.uminho.pt

Vítor Pereira
Centre of Biological Engineering
Department of Informatics
University of Minho, Portugal
LABBELS – Associate Laboratory,
Braga/Guimarães, Portugal
vpereira@ceb.uminho.pt

2022 IEEE Congress on Evolutionary Computation (CEC) | 978-1-6654-6708-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/CEC55065.2022.9870421

***Abstract*—Recent developments in Generative Deep Learning have fostered new engineering methods for protein design. Although deep generative models trained on protein sequence can learn biologically meaningful representations, the design of proteins with optimised properties remains a challenge. We combined deep learning architectures with evolutionary computation to steer the protein generative process towards specific sets of properties to address this problem. The latent space of a Variational Autoencoder is explored by evolutionary algorithms to find the best candidates. A set of single-objective and multi-objective problems were conceived to evaluate the algorithms' capacity to optimise proteins. The optimisation tasks consider the average proteins' hydrophobicity, their solubility and the probability of being generated by a defined functional Hidden Markov Model profile. The results show that Evolutionary Algorithms can achieve good results while allowing for more variability in the design of the experiment, thus resulting in a much greater set of possibly functional novel proteins.**

***Index Terms*—Deep Learning, Generative Models, Protein Design, Evolutionary Algorithms, Novel Proteins**

I. Introduction

Proteins are macromolecules fundamental in organisms, with a wide array of functions within cells, such as catalysing metabolic reactions, providing transport for molecules, or offering structural support and mechanical transduction. The ability to effectively engineer proteins towards macromolecules with the desired traits and functions would result in possibly vast applications [1].

A protein is composed of one or more long chains of amino acids that correspond to the gene's DNA sequence that encodes it. The therapeutic applications resulting from the development of novel proteins, such as enzymes, with enhanced properties and modified applications [2], represent the primary feature currently associated with protein engineering [3]. However, this task is not without its share of obstacles, as the number of potentially functional proteins is minimal within the universe of all potential protein sequences [4].

Despite being fairly recent and relatively unexplored, deep learning generative models have shown the ability to assist protein engineering tasks in the search for novel viable protein structures [5].

De novo protein design has been rapidly developing in the last decade, allowing for the design of a variety of stable novel proteins. Before these successful accomplishments, protein engineering consisted in modifying existing proteins with function and structure similar to the desired [6] [7] [8]. After these initial endeavours, the field of *de novo* protein design evolved exponentially, increasing both the number of published efforts and the computational complexity of the approaches used [7].

Current Deep Learning (DL) approaches have been increasingly used in biological and life sciences thanks to the major steps taken in hardware availability. These approaches to protein design can follow an array of different processes [10]. Mapping a latent space to the sequence space is a common approach usually performed through the use of Autoencoders (AEs), and Generative Adversarial Networks (GANs) [9]. The success of these methods sustains that it is possible to design novel enzymes that carry out completely new reactions. These techniques are more and more used, with applications in the prediction of drug effects or drug repurposing [10].

A. State-of-the-art

Some endeavours have stood out concerning the previous works developed to generate novel proteins applying Variational Autoencoders (VAEs) to the generation. Das et al. proposed a VAE model capable of learning a representation of antimicrobial protein sequences and generating new sequences likely to have antimicrobial properties [11]. Costello and Martin developed BioSeqVAE, a VAE variant that can generate valid protein sequences that are likely to fold and function [1]. Greener et al. used conditional VAEs to design proteins with desired properties and novel protein folds. The authors focused on the generation of metalloproteins [12]. Deep-protein-generation is a framework that uses VAEs designed to work with aligned and unaligned protein sequence data. The objective of this framework is to generate functional variants of luxA bacterial luciferase [13].

B. Goals

This endeavour aims to use VAEs to design new proteins, leveraging the use of EAs to navigate through the latent space. This work's pipeline receives, as input, protein data entries from a supported dataset, carefully prepared to allow the training, validation, and testing of the generative model. The model is validated and evaluated, considering its ability to produce new proteins, targeting desired properties and diversity.

The objective functions used in the present effort are mainly inspired by the methods in Pepfun [14]. This tool implements many methods to study peptides at the structure and sequence level. In the context of this work, the selected evaluation functions were:

- Average hydrophobicity (maximise);
- Rules of synthesis broken (minimise);
- Rules of solubility broken (minimise);
- Maximise the probability of a protein being generated by a Hidden Markov Model (HMM) profile.

1) Hydrophobicity: Hydrophobicity plays a role in the protein's stability, with higher values being correlated to higher stability [15] [16]. This metric is also presumed to be of importance regarding the occurrence of protein aggregation [17] [18]. One objective of optimisation envisaged in this work was to maximise the average hydrophobicity of a protein sequence. To determine this average score, we resorted to the Eisenberg scale [19] for the hydrophobicity of each Amino Acid (AA), averaging the score with basis on the sequence length as Equation 1 presents.

$$\text{Average hydrophobicity} = \frac{\text{sum of AA's hydrophobicity}}{\text{total AAs in the sequence}} \quad (1)$$

2) Synthesis: The original definition for protein synthesis is that the synthesis process represents the translation of the genetic message by the ribosome into a polypeptide [20]. Various Cell-Free Protein Synthesis (CFPS) techniques and machinery have been developed over the last decades to manipulate this process. This process consists in generating the protein through an *in vitro* process as opposed to an *in vivo* one [21].

The events that we aim to avoid and consequently constitute a breach in the rules of synthesis are:

- Two consecutive prolines;
- Presence of the DG and DP motifs;
- Sequence ending with the N or Q AAs;
- Charged residues at every 5 AAs;
- Presence of oxidation-sensitive AAs (M, C or W).

3) Solubility: Protein solubility is a thermodynamic parameter based on the concentration of a protein in a saturated solution in equilibrium with a solid phase. This parameter can be affected by extrinsic factors like pH, temperature, ionic strength, and the presence of solvent additives. Moreover, solubility can also be affected by intrinsic factors, for example, by the AA conformation on the protein surface [22]. The rules of solubility are as follows:

- Discard if the number of charged and/or hydrophobic AAs exceed 45% of the total AAs in the sequence;
- Discard if an absolute total peptide charge with $pH = 7$ greater than +1;
- Discard if more than one glycine or proline in the sequence;
- Discard if the first or the last AAs are charged;
- Discard if an AA represents more than 25% of the whole sequence.

The synthesis and solubility rules used to monitor each protein sequence were extracted from Pepfun [14]. When asserting the rules in the context of an EA, proteins are discarded by giving a fitness of plus infinity.

4) HMM Score: The last objective function maximises the score produced by comparing a generated sequence against an HMM profile. This score is produced through the use of the HMMER software [23].

A profile HMM is a probabilistic model that condenses position-specific information by analysing a multiple sequence alignment and evaluating how conserved each AA is at each position. An HMM allows for the detection of homologs through the existence of significant similarities between protein sequences [24]. The higher this score is, the more likely it is for a generated sequence to be homologous to the sequences used to construct the HMM profile. Therefore, this objective function allows selecting sequences that are more likely to possess the intended functional profile during the EA's evolutionary process and as a post-processing filter. In the particular case of the present work, we retrieved the profile HMM of bacterial luciferase from Pfam [25], a database of HMMs representing an extensive collection of protein families.

5) Net Charge: The net charge at $pH = 7$ is another metric we want to monitor. A protein net charge, at any given pH, is determined by the pK values of the protein ionisable groups [26]. The pK values are a metric of an acid's strength on a base-ten logarithmic scale and are determined by $log10(1/Ka)$ where Ka is an acid dissociation constant [27]. The net charge of a protein is positive at pH values below the Isoelectric point (pI), negative at pH values above the pI and zero at the pI [28]. Although the net charge value of a protein might not represent the most crucial factor in protein stability, the proteins are expected to be less soluble near their respective pI [28]. A higher magnitude in the net charge of the protein should produce an increase or decrease in the pH value, consequently raising the solubility [28].

The literature also indicates that the net charge is an important property of the protein aggregation events [17] [18].

II. Material and Methods

Hawkins-Hooker et al. [13] proposed two different VAE architectures to generate functional variants of the luxA bacterial luciferase. One model was designed to train on raw sequence inputs, while the other on multiple sequence alignments (MSA). The latter, termed MSAVAE, presented the best results and was selected to be used in the present work.

We defined Single-Objective (SOOPs) and Multi-Objective Optimisation Problems (MOOPs) to drive the generative process toward proteins with desired properties. The generative DL model from [13] was trained on a dataset of 69130 homologs of bacterial luciferase, extracted from the InterPro database using the assertion IPR011251. The train/validation split ratio was 80/20.

A. Variational Autoencoders

VAEs increase common autoencoders' capacity by transforming the input's data to statistical distribution parameters instead of compressing it as a fixed code in the latent space. Therefore, VAEs assume that the input data is generated via a statistical process. VAEs randomly select an instance from the distribution through the mean and variance parameters, decoding it into the original input. VAEs are trained via two loss functions: reconstruction and regularisation loss function. The reconstruction loss function ensures that the decoded samples match the original inputs. The regularisation aims at smoothing the latent space created by the encoder by making its distribution resemble a normal one. The latent space's representations are real vectors of size 50.

The schematic representation of the VAE's architecture implemented in this endeavour is presented in Figure 1.

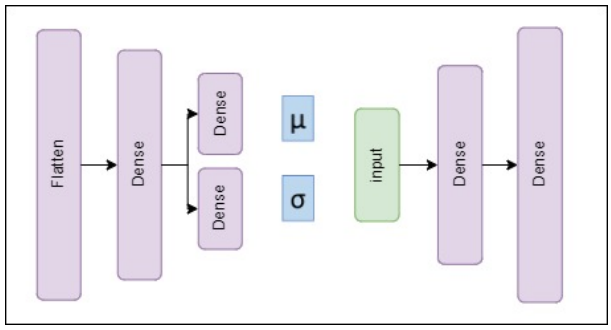

Fig. 1. Schematic representation of the VAE's implemented in this work.

Both the encoder and the decoder use two hidden layers. The latent space is generated using a flatten layer and 2 dense layers. In the decoder, a combination of 2 dense layers was applied. The hidden and output units use, respectively, ReLU and Softmax activation functions.

B. Evolutionary Algorithms

Evolutionary Algorithms (EAs) have proven to be robust methods to solve somewhat complicated optimisation problems [29], allowing the handling and solving of problems with multiple, and even conflicting, objectives in a large and complex search space [30]. Such is the case when conducting protein design optimisation where conflicting goals need to be leveraged to obtain functional proteins. In the context of our experiment, the EAs optimise over the VAE's latent space representations and, therefore, EAs individuals are encoded as real vectors as presented in Figure 2.

.

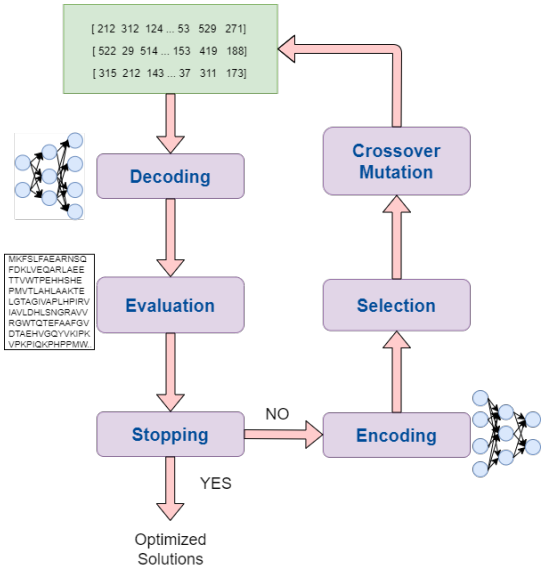

Fig. 2. Schematic representation of the optimization process.

The EAs, implemented using the jMetalPy [31] library, were chosen to accommodate both single and multi-objective optimisation problems, notably, a Genetic Algorithm and three commonly used multi-objective EAs: SPEA2, NSGA-II and NSGA-III. All are configured with a One-Point crossover and two equiprobable mutation operators (a Single and a Gaussian mutator) with respectively 60 and 75% probability of being applied. The EAs use populations of 10 individuals and are run over 160 generations, set as stopping criteria, and 100 repetitions.

C. Code Availability

The experiments use the GenProtEA's framework implementation, developed entirely in Python, which is freely available at https://github.com/martinsM17/GenProtEA. A set of scripts allows building datasets easily by providing an InterPro accession. The protein sequences are automatically trimmed to fit the generative model input and aligned. An HMM protein profile is also produced in cases where none is available from databases such as Pfam.

III. Results and Discussion

We evaluated different sets of proteins taken from the raw dataset, from sampling and decoding the VAE latent space, and from the sets of proteins optimised using the EAs. Each set is composed of 1000 sequences randomly selected. The input of the objective functions is the proteins' amino acid sequence. As such, at each iteration of the EAs, the real vectors, latent space representations, are transformed into amino acid sequences using the VAE's decoder. The VAE's encoder may be used to seed the EA initial population with protein representations known to possess the coveted properties.

A. Single-objective optimizations

We divided the evaluation of the metrics according to the nature of the optimisation problems. The first approach

considers SOOPs leveraging a Genetic Algorithm (GA) that uses a Binary Tournament as a selection mechanism.

The raw dataset's mean net charge value at $pH = 7$ is slightly lesser than the one observed when sampling a large set of proteins from the VAE latent space, notably, and respectively, -8 and -4. When optimising proteins towards a specific goal, the net charge of the final population reflects the objective at hand. When optimising for hydrophobicity, solubility, synthesis, and against the HMM profile, the mean net charge values were -2, -4.8, -1.2 and -8, respectively.

1) Hydrophobicity: In the case of the average hydrophobicity (Figure 3), the distribution of the VAE generated samples presented a tendency toward higher values compared to the ones from the raw dataset.

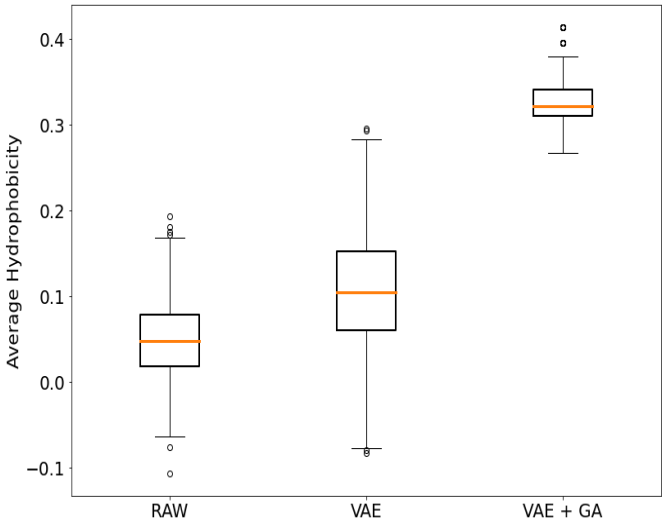

Fig. 3. Comparison between the samples from the raw dataset and samples generated with and without using the GA to optimize the average hydrophobicity. The orange horizontal line represents the median value.

The results show that the protein sequences generated with the VAE architecture have a distribution of much higher values of average hydrophobicity compared to the one presented by the samples extracted from the raw dataset. This comparison indicates that the generated samples showcase not only a higher median value but also a higher Interquartile Range (IQR) and a maximum value of average hydrophobicity.

Optimising the average hydrophobicity using the GA resulted in clear improvements over the previously evaluated samples. Indeed, the median, IQR, maximum and minimum values are higher in GA solution set distribution than in the respective distributions for the raw dataset and direct sampling protein set.

2) Synthesis: Analysing the number of synthesis rules broken (Figure 4), the generated sequences presented much lower values than the ones observed in the distribution of the raw samples. The results point to an overall better distribution of the values concerning the number of synthesis rules broken in the set of generated sequences. Compared with the set of raw samples, the generated sequences, from which proteins were

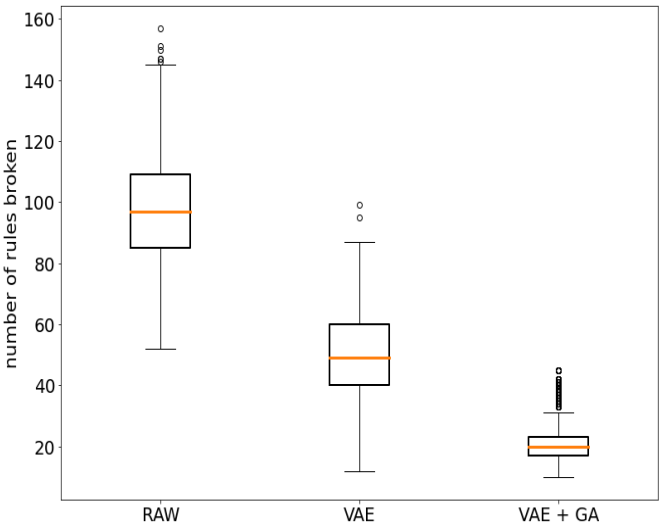

Fig. 4. Comparison between the samples from the raw dataset and samples generated with and without using the GA, concerning the number of synthesis rules broken. The orange horizontal line represents the median value.

taken by direct sampling, present the lowest median value. The same is observed concerning the IQR range, maximum and minimum values. Overall, the generated sequences offer a lower number of synthesis rules broken than the proteins in the raw dataset.

The sequences generated with the use of the GA in the optimisation process present the best results for the case of minimising the synthesis rules broken. The use of the GA resulted in a distribution where the median value, the IQR, and the maximum and minimum values showcase the best results compared with the previously evaluated distributions concerning the number of synthesis rules broken.

3) Solubility: The previous tendency was not carried to the case of the number of broken solubility rules. The results (Figure 5) suggest that the samples from the raw dataset have indeed better outcomes for the case of solubility rules broken when compared with the representatives from the set of generated sequences. It is observable that although the IQR values are similar, the median value for the raw dataset is 2. In contrast, in the case of the generated samples this value is of 3 solubility broken rules, contrasting with the results observed for the Net Charge analysis. Indeed, a similar solubility score would be expected for both datasets, which empathises the need to validate the generated sequences somehow. We next compare both samples using a profile HMM for the raw protein dataset in this context.

All the optimised sequences present the same number of solubility rules broken, 2. Once again, using the GA resulted in clear gains over the previously evaluated distributions. The IQR ranged between 2 and 3, and the overall values ranged from 2 to 4 solubility rules broken.

4) HMM: An additional comparison between the generated and raw samples, the HMM scores, helps elucidate how well the generator can produce new proteins homolog to those it

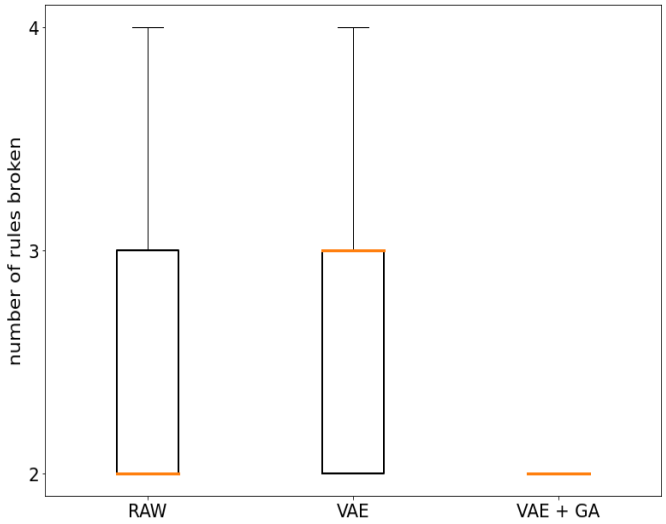

Fig. 5. Comparison between the samples from the raw dataset and samples generated with and without using the GA to optimize the solubility rules broken. The orange horizontal line represents the median value.

was trained on. As expected, the samples from the raw dataset present an overall much better distribution of HMM scores, given that the raw dataset is constructed with luciferase-like oxidoreductases (Figure 6). The best scores from the proteins generated with the VAE architecture are much lower than the ones presented by the raw samples.

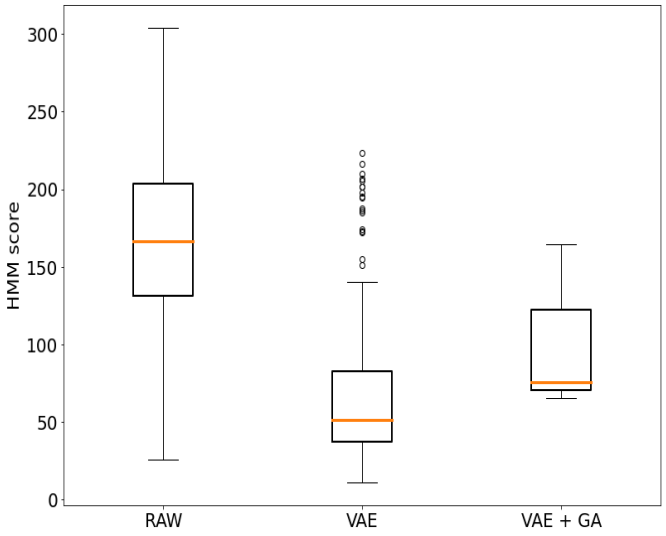

Fig. 6. Comparison between the samples from the raw dataset and samples generated with and without using the GA to optimize the HMM score. The orange horizontal line represents the median value.

This result might indicate that the VAE was unable to capture the distribution of the training set. Nonetheless, one may aim to perform a guided exploration of the latent space, sampling sequences in a more rational way.

The distribution of the samples generated using the GA to maximise the HMM score presents an improvement over the direct sampling of generative model latent space. However, the proteins found by the GA with the best HMM scores have this metric aligned with the median of the raw dataset HMM scores. This result confirms that the VAE could not fully grasp the luxA profile. New generative architectures using distinct representation schemes, such as sequence features (e.g., chemical descriptors and evolutional information) or structural features (e.g., protein surface, secondary structure, and inter-residue distance), are needed.

B. Multi-objective optimizations

One advantage of EAs, when compared with other guided design strategies, is their ability to contemplate the optimisation of more than one objective. As illustrated in the previous section, such is the case in protein design. We defined two MOOPs to evaluate the EAs performance in multi-objective optimisation tasks.

The first consisted in the minimisation of the number of both synthesis and solubility rules broken. For this MOOP we used two different MOEAs: the Non-dominated Sorting Genetic Algorithm-II (NSGA-II) [35] and the Strength Pareto Evolutionary Algorithm 2 (SPEA2) [36]. The second MOOP added a third optimisation objective, the simultaneous maximisation of the HMM score. For this last MOOP we used the Non-dominated Sorting Genetic Algorithm-III (NSGA-III) [37].

1) Net Charge and HMM: The net charge of the samples generated with NSGA-III, optimising the three-objective, present a distribution closer to the raw dataset. The remaining MOEAs resulted in sequences with an overall higher magnitude in the net charge values. This result might indicate a positive correlation between the HMM scores and the net charge values. Indeed, in the single objective optimisation, proteins with optimised HMM had net charge values more aligned to those observed in the training set.

The authors in [13] reported that the vast majority of sequences generated by direct sampling were scored as hits by the HMM (99.7%,) at an E-value threshold of 0.001. We were however, unable to replicate such a result.

2) synthesis and Solubility : NSGA-II and SPEA2 present solutions with similar fitnesses in the minimisation of the number of synthesis (Figure 7) and solubility rules broken (Figure 8). Also, the fitness values achieved by both MOEAs in those objectives are similar to those obtained in single objective optimization.

The number of synthesis rules broken indicates that both SPEA2 and NSGA-II generated samples with good results. These MOEAs present distributions of results with fewer rules broken compared to the other approaches evaluated.

The NSGA-III's solutions presented gains over the direct sampling of the VAE latent space in all but one of the evaluated metrics, the number of synthesis rules broken to which both have similar results. When compared to the other two MOEAs, NSGA-III had a worst performance in the optimisation of solubility and synthesis. Such result highlights the conflicting

nature of optimisation objectives, reinforcing the need for multi-objective optimisations in protein design.

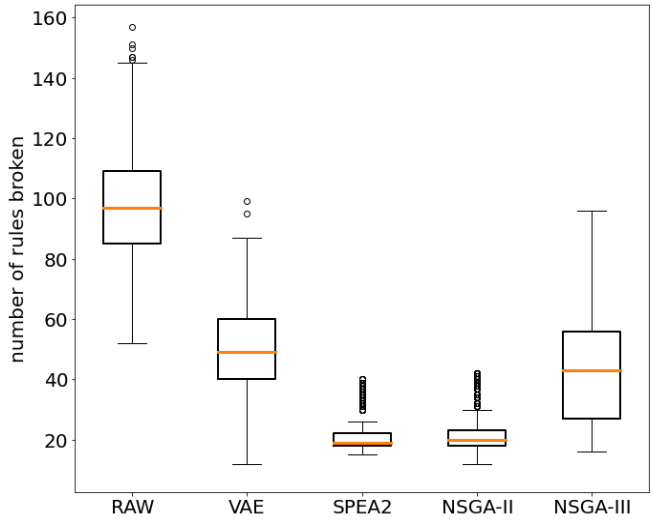

Fig. 7. Comparison of the number of synthesis rules broken between the MOEAs and the previously obtained results.

To analyse the results regarding the number of solubility rules broken, we performed a comparison in the same fashion as the previous one (Figure 8). NSGA-II and SPEA2 present, once again, good results. The majority of the sequences break only 2 rules of solubility, with a few outliers presenting 3 rules broken. The algorithm NSGA-III in its MOOP obtained improvements over the case of generating sequences with VAE and without any EA. The IQR value is the same, but the median value stands at 2 solubility rules broken as opposed to the 3 in the samples generated without an EA in the optimisation.

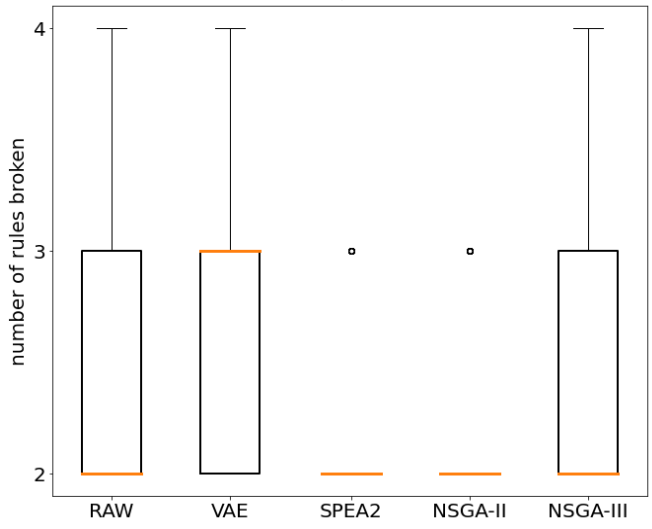

Fig. 8. Comparison of the number of solubility rules broken between the MOEAs and the previously obtained results.

We compared the results obtained by the MOOP using NSGA-III regarding the maximisation of the HMM score with the distributions from the raw dataset and the ones from direct sampling using (Figure 9). The results demonstrate that the generative process using NSGA-III in the optimisation presents an improvement in the overall HMM scores compared to the distribution of the samples generated without applying an EA in the optimisation. These results indicate that the optimisation goal of maximising the HMM score might lead to a better comprehension of the distribution used in the training set.

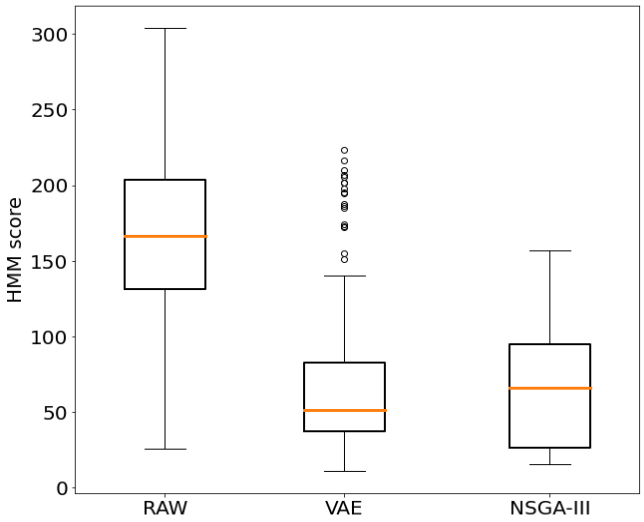

Fig. 9. Comparison of the HMM score between the NSGA-III and the previously obtained results.

C. Discussion

The objective of this work was to construct a generative process capable of generating novel proteins with desired properties using a VAE model and EAs. The optimisation objectives for luxA proteins generation were defined in the basis of the review of the state-of-the-art and literature. However, alternative optimisation goals may be added, such as protein-protein docking in the case of antibodies design.

The state-of-the-art review did not reveal any benchmarks for the optimisation problems in the computational generation of novel protein sequences with desirable properties. Since an approach like the one developed in this work was not found in any of the literature reviewed, the optimisation problems were defined according to the literature on protein design, specifically regarding protein sequence manipulation.

The generative process resulted in distributions consistently different from the ones observed in the original datasets. The protein designs guided by EAs optimisation resulted in better solutions than those collected by direct sampling. Furthermore, the MOEAs allow the selection of proteins that show evidence of having the best set of properties.

As previously mentioned, the model used in this work was the one with the best results in the work of Hawkins-Hooker et al. [13]. In their assessment, the authors refer that most of the

proteins experimentally tested showcased high luminescence levels (a key feature of the proteins in the training dataset). The evaluations performed in our endeavour indicated that, in terms of the evaluation functions used, the model alone could not capture the distributions observed in the original dataset. The use of EAs, introduced in our work, offers a way to mitigate this problem and might result in an even higher success rate of the generated proteins.

D. Protein Structure Prediction

To infer the protein structure generated through our designed experiments, we first selected the best-generated sample from the case study using the NSGA-III. The chosen protein had the highest cumulative score resulting from subtracting the number of synthesis rules broken from the HMM score. The resulting sequence was:

```
MKFSLFAEARNSQTHRFDKLVEQARLAEERDFTTVWTPEHHSHEFSPSPMVT
LAHLAAKTERVALGTAGIVAPLHPIRVAKEIAVLDHLSNGRAVVGFARGWTQ
TEFAAFGVASRQAGLREIVDAIQKLWADDTAEHVGQYVKIPKATAVPKPIQK
PHPPMWVAQGGPENFKWAAEHGAGFMVTLLGGLEEIEKRIKEFREAFDHEDP
KVAVLRHTHVTNKDGVRNVAIQFKREFSVQKNRRAEIAELADFTDESFHKRG
VFGSVDEVVDRLERLDGVDEIALDAKEVLDGLALLQEQHRAYFRA
```

In this case, the selected sample presents 57 synthesis rules broken and an HMM score of 156,3. After selecting the protein sequence, we leveraged the *AlphaFold* capabilities to infer the protein's structure. The *AlphaFold* software is a computational tool developed by DeepMind [32] that predicts a protein's 3D structure from its amino acid sequence [33] [34].

The LDDT measures the percentage of correctly predicted interatomic distances and the pLDDT indicates the confidence in the local structure. The pLDDT metric ranges from 0 to 100 (Figure 10), and the presented predicted structure (Figure 11) showcases the levels of metrics in a colour-coded fashion.

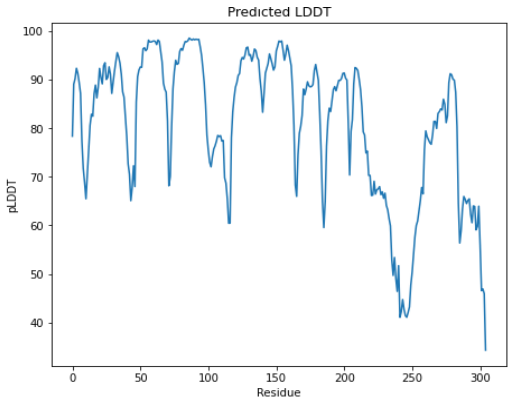

Fig. 10. Predicted per-residue estimate confidence on a scale from 0 to 100.

The results showcase a majority of the residues scored with *very high* or confident levels of confidence according to pLDDT values.

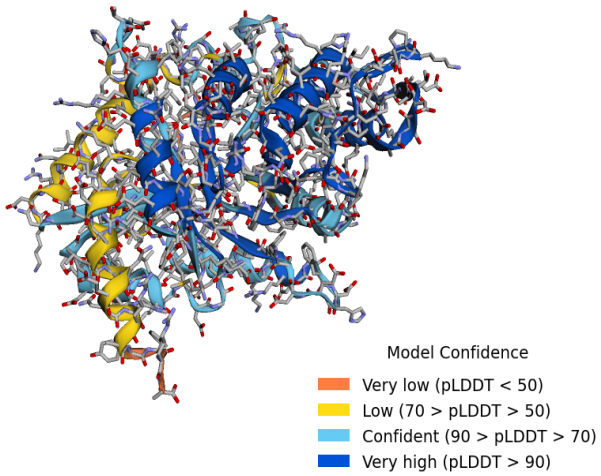

Fig. 11. AlphaFold's predicted structure of the selected protein sequence.

To determine AlphaFold's confidence in the relative positions, another metric must be computed, the predicted aligned error (Figure 12).

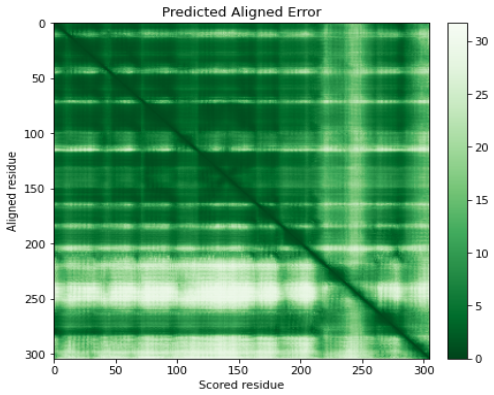

Fig. 12. AlphaFold's predicted aligned error of the selected protein sequence.

The higher the error associated with the residue's position, the more faded tone of green. The results of the predicted aligned error indicate, as in the case of the pLDDT metric, a high level of confidence in the relative predicted positions of each residue, with some shallow levels of confidence on some residues.

E. Similarity Search

The protein sequence selected for the structure prediction task was put through a similarity search using the protein Blast tool. The best result is of a LLM class flavin-dependent oxidoreductase with a total and maximum scores of 200, a query cover of 94% and an *E value* of $4e^{-58}$ and a per cent identity of 41.28%.

IV. Conclusion

Solubility, synthesis, stability, and aggregation of proteins are vital aspects of protein design. We demonstrated that it is possible to optimise these properties by combining deep generative models and evolutionary computation to generate and explore latent spaces of protein embeddings.

The developed work sets a foundation for future endeavours opening a path to a wide range of further applications to this process. There are various ways to improve the obtained results by, for example, tuning the implemented architectures using transfer learning, using new generative model architectures (e.g., Geometric Convolutional VAE), or designing further case studies using different EAs and objective functions (e.g. docking). In particular, new biological problems may be addressed by exploring the presented architecture, for example, the design of antibodies.

Aknowlegments

This project has received funding from the European Union's Horizon 2020 research and innovation programme (Grant Agreement Number 814408).

References

[1] Zak Costello and Hector Garcia Martin. How to hallucinate functional proteins. arXiv preprint arXiv:1903.00458, 2019 Unpublished.

[2] H.A. Daniel Lagassé, Aikaterini Alexaki, Vijaya L. Simhadri, Nobuko H. Katagiri, Wojciech Jankowski, Zuben E. Sauna, and Chava Kimchi-Sarfaty. Recent advances in (therapeutic protein) drug development. F1000Research, 6:113, feb 2017.

[3] Ole Kirk, Torben Vedel Borchert, and Claus Crone Fuglsang. Industrial enzyme applications. Current Opinion in Biotechnology, 13(4):345–351, aug 2002.

[4] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. arXiv preprint arXiv:1712.03346, 2017 Unpublished.

[5] Namrata Anand and Po-Ssu Huang. Generative modeling for protein structures. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 7505–7516, 2018

[6] Enrique Marcos and Daniel-Adriano Silva. Essentials of de novo protein design: Methods and applications. Wiley Interdisciplinary Reviews: Computational Molecular Science, 8(6):e1374, 2018.

[7] Ivan V Korendovych and William F DeGrado. De novo protein design, a retrospective. Quarterly reviews of biophysics, 53, 2020.

[8] Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. Nature, 537(7620):320–327, 2016.

[9] Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J Gray. Deep learning in protein structural modeling and design. Patterns, page 100142, 2020.

[10] Antonella Paladino, Filippo Marchetti, Silvia Rinaldi, and Giorgio Colombo. Protein design: from computer models to artificial intelligence. Wiley Interdisciplinary Reviews: Computational Molecular Science, 7(5):e1318, 2017

[11] Payel Das, Kahini Wadhawan, Oscar Chang, Tom Sercu, Cicero Dos Santos, Matthew Riemer, Vijil Chenthamarakshan, Inkit Padhi, and Aleksandra Mojsilovic. Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences. arXiv preprint arXiv:1810.07743, 2018 Unpublished.

[12] Joe G Greener, Lewis Moffat, and David T Jones. Design of metalloproteins and novel protein folds using variational autoencoders. Scientific reports, 8(1):1–12, 2018.

[13] Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. PLoS computational biology, 17(2):e1008736, 2021

[14] Rodrigo Ochoa and Pilar Cossio. Pepfun: Open source protocols for peptide-related computational analysis. Molecules, 26(6):1664, 2021.

[15] Shambhu Malleshappa Gowder, Jhinuk Chatterjee, Tanusree Chaudhuri, and Kusum Paul. Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. The Scientific World Journal, 2014, 2014.

[16] Caroline Strub, Carole Alies, Andrée Lougarre, Caroline Ladurantie, Jerzy Czaplicki, and Didier Fournier. Mutation of exposed hydrophobic amino acids to arginine to increase protein stability. BMC biochemistry, 5(1):1–6, 2004.

[17] Martino Calamai, Niccolo Taddei, Massimo Stefani, Giampietro Ramponi, and Fabrizio Chiti. Relative influence of hydrophobicity and net charge in the aggregation of two homologous proteins. Biochemistry, 42(51):15078–15083, 2003.

[18] Fabrizio Chiti. Relative importance of hydrophobicity, net charge, and secondary structure propensities in protein aggregation. In protein misfolding, aggregation, and conformational diseases, pages 43–59. Springer, 2006.

[19] David Eisenberg, Robert M Weiss, and Thomas C Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. Proceedings of the National Academy of Sciences, 81(1):140–144, 1984.

[20] CG Carter and DF Houlihan. Protein synthesis. Fish physiology, 20:31–75, 2001.

[21] Erik D Carlson, Rui Gan, C Eric Hodgman, and Michael C Jewett. Cell-free protein synthesis: applications come of age. Biotechnology advances, 30(5):1185–1194, 2012.

[22] Ryan M Kramer, Varad R Shende, Nicole Motl, C Nick Pace, and J Martin Scholtz. Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. Biophysical journal, 102(8):1907–1915, 2012.

[23] Finn, Robert D., Jody Clements, and Sean R. Eddy. "HMMER web server: interactive sequence similarity searching." Nucleic acids research 39.suppl_2 (2011): W29-W37.

[24] Sean R. Eddy. Profile hidden markov models. Bioinformatics (Oxford, England), 14(9):755–763, 1998.

[25] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, et al. The pfam protein families database in 2019. Nucleic acids research, 47(D1):D427–D432, 2019.

[26] Charles Tanford. The interpretation of hydrogen ion titration curves of proteins. Advances in protein chemistry, 17:69–165, 1963.

[27] Antonio Sillero and João Meireles Ribeiro. Isoelectric points of proteins: theoretical determination. Analytical biochemistry, 179(2):319–325, 1989.

[28] Kevin L Shaw, Gerald R Grimsley, Gennady I Yakovlev, Alexander A Makarov, and C Nick Pace. The effect of net charge on the solubility, activity, and stability of ribonuclease sa. Protein Science, 10(6):1206–1215, 2001.

[29] Ali Wagdy Mohamed. Solving large-scale global optimisation problems using enhanced adaptive differential evolution algorithm. Complex & Intelligent Systems, 3(4):205–231, 2017.

[30] David A Van Veldhuizen and Gary B Lamont. On measuring multiobjective evolutionary algorithm performance. In Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No. 00TH8512), volume 1, pages 204–211. IEEE, 2000.

[31] Antonio Benítez-Hidalgo, Antonio J Nebro, José García-Nieto, Izaskun Oregi, and Javier Del Ser. jmetalpy: A python framework for multi-objective optimization with metaheuristics. Swarm and Evolutionary Computation, 51:100598, 2019.

[32] Powles, Julia, and Hal Hodson. "Google DeepMind and healthcare in an age of algorithms." Health and technology 7.4 (2017): 351-367.

[33] Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.

[34] Varadi, Mihaly, et al. "AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models." Nucleic acids research 50.D1 (2022): D439-D444.

[35] Deb, Kalyanmoy, et al. "A fast and elitist multi-objective genetic algorithm: NSGA-II." IEEE transactions on evolutionary computation 6.2 (2002): 182-197.

[36] Zitzler, Eckart, Marco Laumanns, and Lothar Thiele. "SPEA2: Improving the strength Pareto evolutionary algorithm." TIK-report 103 (2001).

[37] Kalyanmoy Deb and Himanshu Jain. An evolutionary many-objective optimisation algorithm using reference-point-based nondominated sorting approach, part i: solving problems with box constraints. IEEE transactions on evolutionary computation, 18(4):577– 601, 2013.