From Sequence to Stability: Rational Insulin Design via Genetic Algorithms and Deep Learning Models in Structural Bioinformatics

Mateus R. Gomes[1], Gabriel Vasconcelos Fruet[1], Ingryd Medeiros[2], Roner F. da Costa[2], Eveline M. Bezerra[2], Danielo G. Gomes[1]

[1]Bacharelado em Engenharia de Computação
Grupo de Redes, Engenharia de Software e Sistemas (GREat)
Centro de Tecnologia - Universidade Federal do Ceará (UFC)

[2]Programa de Pós-Graduação em Ciência e Engenharia de Materias - PPgCEM
Laboratório de Física Computacional (LFC)
CCEN – Universidade Federal Rural do Semi-Árido (UFERSA)

[matribg04,gabrielfruet]@alu.ufc.br

ingryd.medeiros@alunos.ufersa.edu.br

[roner.costa,eveline.bezerra]@ufersa.edu.br

danielo@ufc.br

Abstract. *Insulin's therapeutic efficacy is hindered by its thermal instability, a major limitation in global diabetes treatment, especially in regions lacking reliable refrigeration. Here we present an in silico protein design pipeline that integrates a multi-objective genetic algorithm with deep learning models to engineer insulin variants with enhanced thermostability and reduced aggregation propensity. The algorithm evolves populations of mutated insulin sequences, evaluated by TemBERTure for thermostability and Aggrescan3D for solubility, incorporating ESMFold for 3D structure prediction. A microservices architecture using Docker ensures scalable and efficient execution. Our results identify candidate variants that maintain high sequence identity and preserve key functional motifs while showing superior biophysical properties. These findings illustrate how the combination of evolutionary algorithms and protein language models can support rational, data-driven strategies in protein engineering. The source code and reproducible experiments are publicly available at https://github.com/gabrielfruet/protein-aggregation.*

1. Introduction

Diabetes remains a major global health challenge, affecting over 415 million adults in 2015, with projections exceeding 640 million by 2040 [Ogurtsova et al. 2017]. Since the discovery of insulin in 1921, its therapeutic use has revolutionized diabetes care, with the first successful human treatment earning Banting and colleagues the Nobel Prize in 1923 [Polonsky 2012] [Sanger 1959].

Insulin is a peptide hormone essential for glucose homeostasis. Secreted by pancreatic β-cells, it binds to insulin receptors on target cells, triggering a signaling cascade

that facilitates glucose uptake via GLUT transporters [Polonsky 2012]. However, insulin is thermolabile: it loses stability above around 30°C, leading to structural changes that impair receptor binding and biological function. This thermosensitivity presents significant challenges for storage and transport, particularly in low-resource settings where refrigeration is unreliable. Improving the thermal stability of insulin is therefore a priority for global health, with potential to enhance treatment access and equity.

Recent advances in computational biology offer powerful tools for rational protein design. Genetic algorithms (GAs) [Holland 1992], inspired by evolutionary processes, provide an efficient framework for exploring complex mutational landscapes. In protein engineering, GAs can identify amino acid substitutions that enhance stability without compromising function.

Simultaneously, protein language models (pLMs) [Elnaggar et al. 2022], based on transformer architectures, have demonstrated strong performance across sequence-based prediction tasks. By interpreting protein sequences as linguistic constructs, pLMs capture the implicit grammar that governs folding and function.

Structure prediction tools such as ESMFold further extend these capabilities, enabling high-confidence 3D structure inference from sequence alone. These models facilitate downstream analyses of solubility, aggregation, and stability—traits not readily inferred from sequence alone.

Here, we present a computational pipeline that integrates a genetic algorithm with structural prediction to evolve insulin variants with improved thermostability and reduced aggregation propensity. A multi-objective fitness function was designed to guide this evolution, combining predictive scores for stability and solubility. This study illustrates the potential of combining AI-driven methods with evolutionary search to design biophysically enhanced protein therapeutics.

2. Materials and Methods

Human insulin (molecular formula $C_{257}H_{383}N_{65}O_{77}S_6$, PubChem CID 118984375) is composed of 51 amino acids, arranged in two polypeptide chains: the a-chain (21 amino acids) and the b-chain (30 amino acids), linked by two interchain disulfide bonds (a7–b7 and a20–b19), and stabilized further by one intrachain disulfide bond within the a-chain (a6–a11) [Sanger 1959]. The primary sequence of the A-chain is: GIVEQCCTSICSLYQLENYCN, while the B-chain is: FVNQHLCGSHLVEALYLVCGERGFFYTPKT. This precise arrangement is critical for insulin's three-dimensional conformation, which in turn governs its ability to bind effectively to the insulin receptor and trigger downstream signaling. Any alteration in the amino acid sequence can affect the folding, receptor affinity, and even the rate of aggregation or degradation of the molecule. Thus, understanding the relationship between sequence, structure, and function is essential for the rational design of insulin analogs with improved pharmacokinetic and pharmacodynamic properties.

2.1. Genetic Algorithm Framework

The genetic algorithm was implemented using the *PyGAD* Python library (v3.4.0). To generate the initial population, we used a seed sequence from a common insulin therapy analog, which consists of the concatenated A and B chains: GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKA. Each individual was

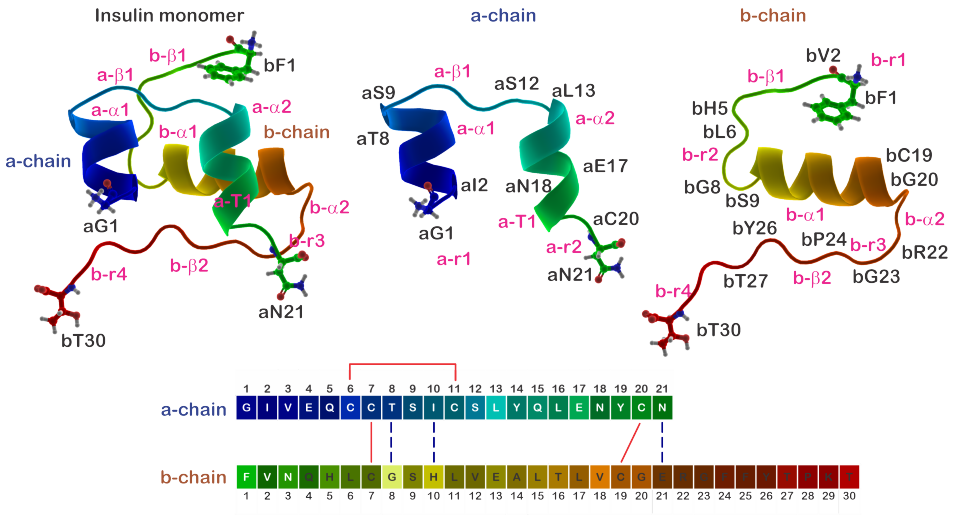

Figure 1. The insulin monomer (left), its a-chain (center), and b-chain (right), along with their labeled residues (black letters). The secondary structures of insulin are represented as follows: α-helices, β-sheets, turns, and random coils, colored in blue/green (for the a-chain) and green/red (for the b-chain), respectively. These structures are labeled in pink as a-chain (b-chain): a-r1, a-α1, a-β1, a-α2, a-T1 and a-r2 (b-r1, b-β1, b-r2, b-α1, b-α2, b-r3, b-β2, and b-r4). The primary sequence of the insulin a-chain (21 residues) and b-chain (30 residues) is shown below, colored according to the corresponding structural regions. The intra-chain disulfide bond (aC6–aC11) and inter-chain disulfide bonds (aC7–bC7 and aC20–bC19) are indicated with solid red lines. Dashed blue lines represent strong noncovalent interactions between key residues. PDB ID 2JV1 [Bocian et al. 2008].

created by copying this parent sequence and introducing a single random point mutation at a chosen position. Each amino acid sequence was converted into its numerical encoding for subsequent analyses.

Evolution was carried out over twenty generations with a population size of 64 and 32 parents selected per generation initially. Genetic operators (crossover and mutation) were applied to produce offspring, the fitness was defined as a weighted composite function based on the predicted thermostability and aggregation propensity. At the end of each generation, the entire population (together with its metrics) was recorded for further review.

2.2. Multi-Objective Fitness Function

To compute the first term of our fitness function, we used TemBERTure [Rodella et al. 2024], a deep-learning framework for protein thermostability prediction. In particular, we employed the TemBERTureCLS model, a ProtBERT [Elnaggar et al. 2022] fine-tuned classification model, that receives FASTA-formatted sequences as inputs and produces a thermophilicity score $T \, \epsilon \, [0, 1]$ with an overall accuracy of 0.89.

For the second metric, three-dimensional structures of each variant were generated via ESMFold [Rives et al. 2021], a transformer-based neural network developed by Meta. Based on the PDB structures, we employed the AGGRESCAN3D 2.0

[Zambrano et al. 2015] server to evaluate their aggregation propensity and solubility. The server accepts one or more PDB files as input and returns four metrics:

1. Minimal score value: Value of the most soluble residue in the structural context.
2. Maximal score value: Value of the most aggregation-prone residue in the structural context.
3. Average score: A normalized indicator of the aggregation propensity/solubility of the protein structure. Allows comparing the solubility of different protein structures. It also allows assessing changes in solubility promoted by amino acid substitutions in a particular protein structure. The more negative the value, the highest the normalized solubility.
4. Total score: A global indicator of the aggregation propensity/solubility of the protein structure. It depends on the protein size. It allows assessing changes in solubility promoted by amino acid substitutions in a particular protein structure. The more negative the value, the highest the global solubility.

The average score (3) was selected as it best aligned with the objectives of our study. To incorporate it efficiently into the fitness function, we performed a relative normalization at each generation of the genetic algorithm as follows:

$$A_{\text{norm}} = 1 - \frac{A_i - A_{\text{min}}}{A_{\text{max}} - A_{\text{min}}} \tag{1}$$

where:

- A_i is the aggregation-propensity score for a given individual.
- A_{min} is the minimum score within the current generation.
- A_{max} is the maximum score within the current generation.

This transformation truncates the aggregation-propensity values to the interval $[0, 1]$ and inverts the scale by subtracting from 1, so that lower propensity scores (i.e. more soluble variants) result in higher normalized values.

Our final fitness function was defined by assigning a weight of 60% to the thermostability score and 40% to the normalized aggregation score.

$$f(T, A_{\text{norm}}) = 0.6 \cdot T + 0.4 \cdot A_{\text{norm}} \tag{2}$$

2.3. Pipeline and Architecture

The computational pipeline was architected using a microservices approach, with each component encapsulated within Docker containers to resolve dependency conflicts between the distinct PyTorch versions required by TemBERTure and ESMFold. Specifically, two containers were deployed:

1. Core Service Container: Orchestrates the genetic algorithm and performs local thermostability predictions using the TemBERTure model.
2. ESMFold Service Container: Exposes the ESMFold model via a RESTful API, decoupling its execution environment from the core pipeline.

Inter-service communications and external calls to the AGGRESCAN3D server were optimized to minimize per-generation computational overhead. We adopted an asynchronous processing model, enabling multiple PDB retrievals and aggregation-propensity scoring requests to be handled in batch. A caching layer was also introduced to avoid redundant PDB computations for sequences that had already been evaluated. For the AGGRESCAN3D RESTful API requests, counting semaphores were employed to regulate concurrency and prevent blocking due to excessive requests. Figure 2 visually represents the archicterure and pipeline.

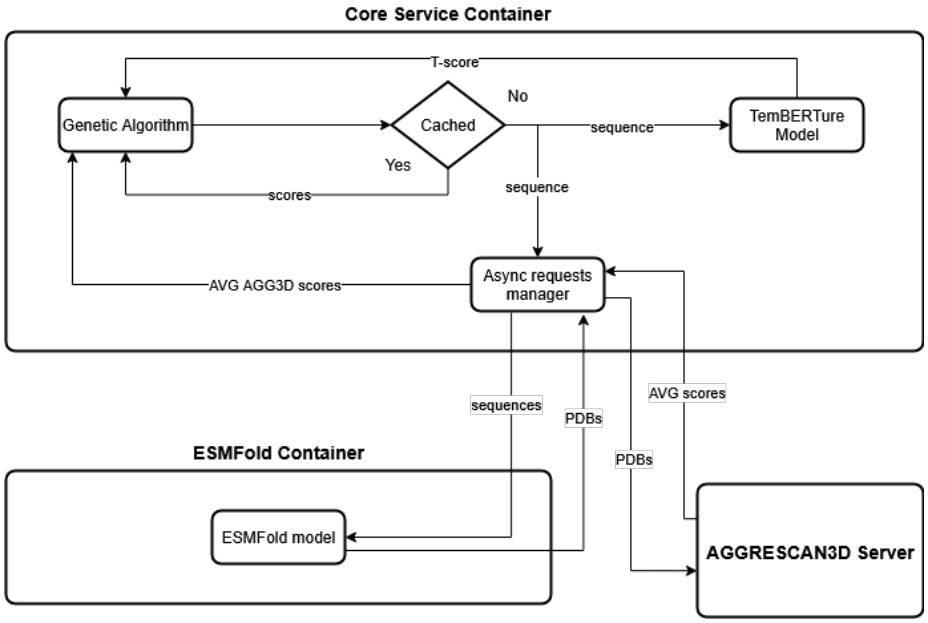

Figure 2. Architecture flowchart.

3. Results

Table 1 presents the percentage identity of each generation's candidate sequences relative to the seed insulin seed molecule. Generations 5 and 6 emerge as the most promising, as both achieve an identity of 94.1% in their closest individual variant and maintain average identities between 80% and 85%. This performance reflects a balance between exploration and conservation. These populations introduce only minimal substitutions outside of functionally critical residues, while preserving the cysteine amino acids, which are essential for insulin bioactivity.

It is also possible to infer from Figure 3 that between generations 6 and 9 a plateau in fitness at 0.511 was reached. This behavior motivated us to modify certain parameters of the genetic algorithm, namely the number of selected parents from 32 to 20, and the mutation probability from 1/51 to 2/51 during the evolutions. With these modifications, the molecules demonstrated increased fitness but decay in the biological function of the insulin molecule, losing essential characteristics, such as critical cysteines. This confirms the rapid decay in the measurements in Table 1 from generation 9 onward.

The violin plot in Figure 4 illustrates that the mean fitness scores across the initial generations were maintained between 0.15 and 0.4. However, after the tenth generation,

Generation	Best (%)	Average (%)	Generation	Best (%)	Average (%)
0	100.0	96.1	**10**	80.4	70.2
1	100.0	95.5	**11**	80.4	68.4
2	100.0	90.8	**12**	74.5	65.8
3	98.0	89.5	**13**	74.5	65.5
4	94.1	86.6	**14**	76.5	62.3
5	94.1	84.5	**15**	70.6	56.9
6	94.1	81.9	**16**	62.7	55.3
7	86.3	79.1	**17**	62.7	52.4
8	88.2	75.7	**18**	62.7	51.1
9	86.3	73.6	**19**	58.8	48.2
			20	54.9	46.8

Table 1. Generational amino acids percentage identity between evolved sequences and the seed insulin molecule.

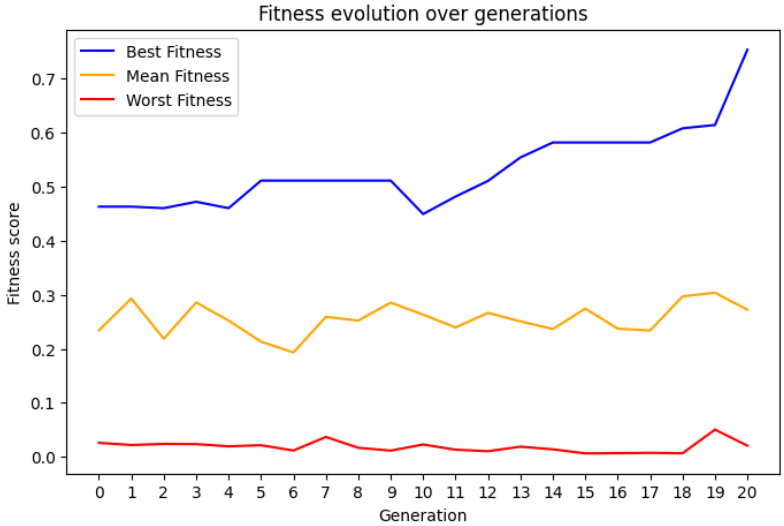

Figure 3. Fitness value evolution over generations.

following a modification of the evolutionary parameters to address stagnation at generation nine, the subsequent increase in fitness scores demonstrates that the enhanced exploratory pressure successfully allowed the algorithm to escape a local optimum. This gain was achieved at the cost of the molecules' functional bio-activity, leading to the loss of critical structural integrity.

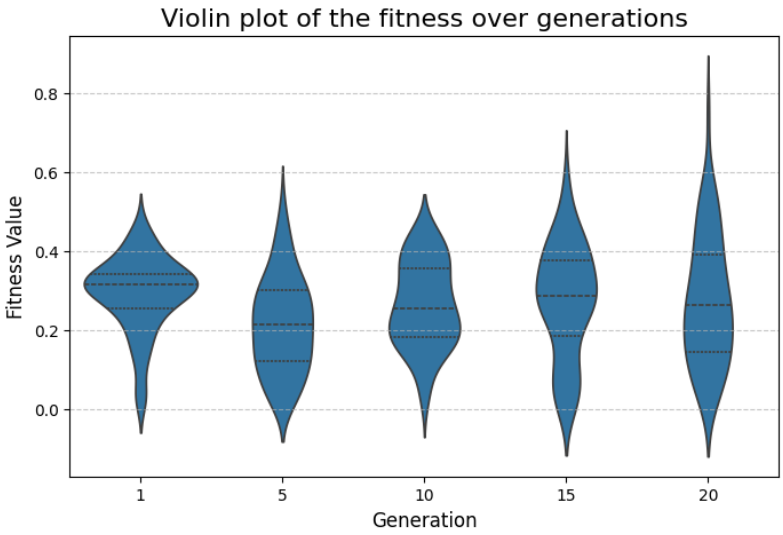

Figure 4. Violin plot of the fitness over generations.

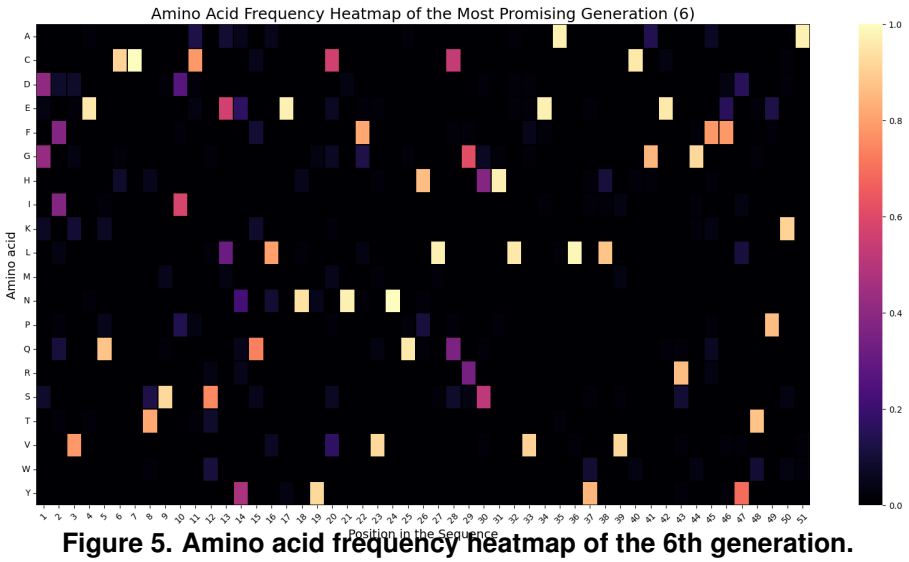

Figure 5. Amino acid frequency heatmap of the 6th generation.

The analysis of the amino acid frequency matrix from the selected generation (Figure 5) reveals a high degree of conservation among the key residues that characterize the insulin molecule. This confirms that the algorithm is not performing a blind or random search, but is effectively learning and preserving the structural patterns essential to the protein's architecture.

Some molecules have demonstrated promise, exhibiting enhanced thermostability and a reduced propensity for aggregate formation compared to the base insulin, while retaining a high percentage of molecular proximity to the seed and also preserving critical regions.

Among the populations, there were two main promising candidates. The first being **DFVEQCCTSICSAYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYT** (A and B sequences concatenated), from generation 6 prioritizes bioactivity conservation, with a thermostability score of 0.14 and an aggregation propensity of -0.32

(not normalized) while preserving 94.1% sequence identity to seed insulin and retaining all six canonical cysteines as well as the core b-chain receptor-binding motif. The second one, **DYVELCCHSICSENQLENYCNFVNQHACGSCAVEALYLYCAERGKFYTEYA** (A and B sequences concatenated), from generation 12, represents a thermostability-enhanced variant, achieving a thermostability score of 0.24 and an aggregation propensity of -0.67 (not normalized) while maintaining 72.6% identity to the seed molecule. This sequence introduces targeted substitutions in both its N-terminal and C-terminal regions, which could compromise receptor-binding contacts.

For comparative purposes, the seed molecule used achieved a thermostability score of 0.05 and an aggregation propensity score of 0.05.

The structural configurations of insulin were analyzed starting from its crystallographic structure, to investigate conformational changes across different generations of optimization. The three-dimensional structures of each variant were generated using ESMFold and PyMOL [Schrödinger, LLC 2015]. Figure 6 presents a cartoon representation of the 3D configurations of the insulin molecule: (A) seed molecule; (B) the candidate structure from the 6th generation; and (C) the candidate structure from the 12th generation.

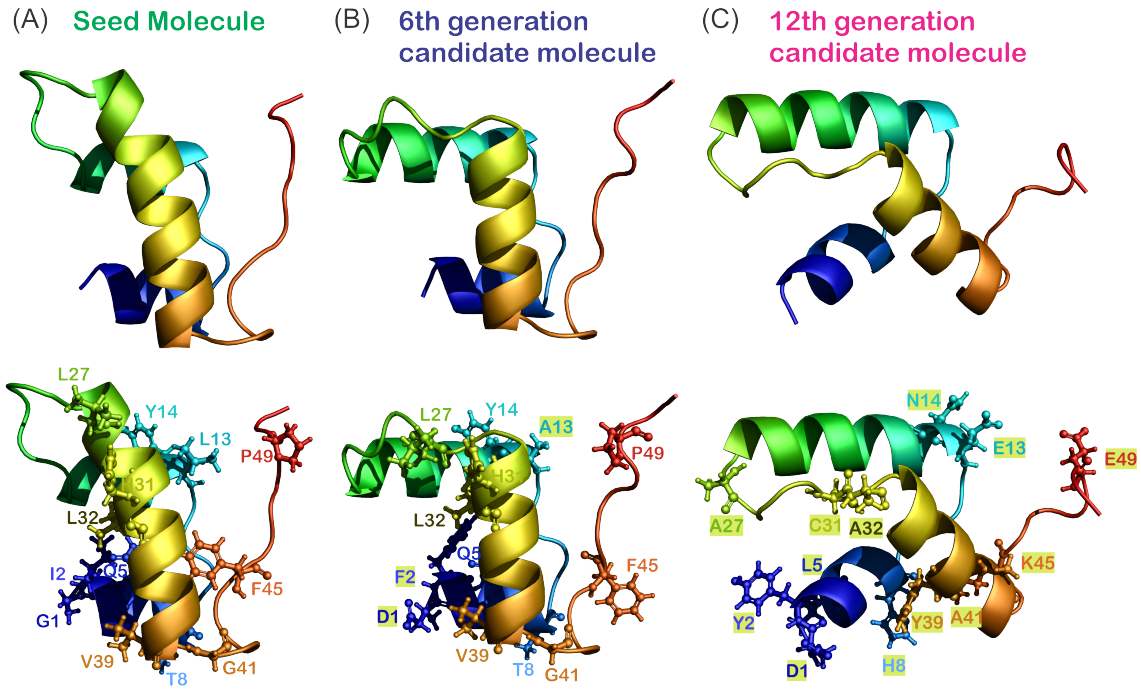

Figure 6. Three-dimensional representations of insulin molecular conformations throughout a computational evolutionary process are presented as follows: (A) The seed molecule serves as the starting point. (B) The 6th generation candidate molecule exhibits specific residue mutations while preserving the overall protein structure. (C) The 12th generation candidate molecule features multiple amino acid substitutions, indicating further structural optimization. The bottom views emphasize key residues that contribute to structural stability, with color-coded chains and mutated residues highlighted in yellow boxes.

4. Conclusion

Here we propose a genetic algorithm to design insulin variants with improved thermostability and reduced aggregation propensity. The fitness function combined thermostability prediction scores with Aggrescan3D metrics, enabling a more comprehensive evaluation. Structural models generated by ESMFold were essential for assessing three-dimensional features, particularly solubility and aggregation risk.

Our *in silico* evolution strategy successfully produced insulin variants with higher predicted thermostability and lower aggregation scores. Variants from generations 5 and 6 were especially promising, as they retained the sequence identity and bioactive features of native human insulin. This suggests an effective balance between structural conservation and optimization. In contrast, later generations, despite showing improved fitness, lost sequence identity—indicating potential loss of critical molecular properties. These findings highlight the potential of genetic algorithms in designing more stable insulin variants, as long as bioactivity is preserved.

It is important to note that these results are based on *in silico* simulations. While computational models provide valuable predictions, experimental validation is essential to confirm thermostability and bioactivity *in vitro*.

Overall, this paper lays the groundwork for future studies involving structural energy-based classification and molecular dynamics simulations. These steps will help evaluate the stability of selected variants under different temperature conditions and identify the most promising candidates for further development.

Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil (CAPES) - Finance Code 001. Danielo G. Gomes and Mateus R. Gomes acknowledge the support from the National Council for Scientific and Technological Development (CNPq), Brazil (grant nos. 311845/2022-3 and 110386/2025-6).

References

[Bocian et al. 2008] Bocian, W., Sitkowski, J., Bednarek, E., Tarnowska, A., Kawecki, R., and Kozerski, L. (2008). Structure of human insulin monomer in water/acetonitrile solution. *Journal of Biomolecular NMR*, 40:55–64.

[Elnaggar et al. 2022] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2022). Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127.

[Holland 1992] Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, Cambridge, MA, USA.

[Ogurtsova et al. 2017] Ogurtsova, K., da Rocha Fernandes, J., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N., Cavan, D., Shaw, J., and Makaroff, L. (2017). Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, 128:40–50.

[Polonsky 2012] Polonsky, K. S. (2012). The past 200 years in diabetes. *New England Journal of Medicine*, 367:1332–1340.

[Rives et al. 2021] Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118. bioRxiv 10.1101/622803.

[Rodella et al. 2024] Rodella, C., Lazaridi, S., and Lemmin, T. (2024). TemBERTure: advancing protein thermostability prediction with deep learning and attention mechanisms. *Bioinformatics Advances*, 4(1):vbae103.

[Sanger 1959] Sanger, F. (1959). Chemistry of insulin. *Science*, 129:1340–1344.

[Schrödinger, LLC 2015] Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8.

[Zambrano et al. 2015] Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., and Ventura, S. (2015). Aggrescan3d (a3d): server for prediction of aggregation properties of protein structures. *Nucleic Acids Research*, 43(W1):W306–W313.