TMP-M2Align: A Topology-Aware Multiobjective Approach to the Multiple Sequence Alignment of Transmembrane Proteins

Joel Cedeño-Muñoz , Cristian Zambrano-Vega , Antonio J. Nebro *

Posted Date: 2 October 2025

doi: 10.20944/preprints202510.0109.v1

Keywords: multiple sequence alignment; transmembrane proteins; multiobjective optimization; evolutionary algorithms; GPCR

 doi:10.20944/preprints202510.0109.v1

Article

TMP-M2Align: A Topology-Aware Multiobjective Approach to the Multiple Sequence Alignment of Transmembrane Proteins

Joel Cedeño-Muñoz [1], Cristian Zambrano-Vega [2] and Antonio J. Nebro [3,4]*

[1] Faculty of Animal and Biology Sciences, State Technical University of Quevedo, Quevedo, 120503, Ecuador
[2] Faculty of Computer Science and Digital Design, State Technical University of Quevedo, Quevedo, 120503, Ecuador
[3] ITIS Software, Edificio de Investigación Ada Byron, University of Málaga, Málaga, 29071, Spain
[4] Dept. de Lenguajes y Ciencias de la Computación, University of Málaga, Málaga, 29071, Spain
* Correspondence: ajnebro@uma.es

Abstract

Transmembrane proteins (TMPs) constitute approximately 30% of the mammalian proteome and are critical targets in biomedical research due to their involvement in signaling, transport, and drug interactions. However, their unique structural characteristics pose significant challenges for conventional multiple sequence alignment (MSA) methods, which are typically optimized for soluble proteins. In this paper, we propose TMP-M2Align, a novel topology-aware multiobjective algorithm specifically designed for the multiple alignment of TMPs. The method simultaneously optimizes two complementary objectives: (i) a topology-aware Sum-of-Pairs (SP) score that integrates region-specific substitution matrices and gap penalties, and (ii) an Aligned Regions (AR) score that rewards consistent alignment of functional and topological domains. By combining these objectives, TMP-M2Align generates Pareto front approximations of alignment solutions, enabling researchers to select trade-offs that best suit their biological questions. We evaluated TMP-M2Align on BAliBASE Reference Set 7 and on complete datasets of human G protein-coupled receptors (GPCRs) from classes A, B1, and C. Experimental results demonstrate that TMP-M2Align consistently outperforms both traditional alignment tools and specialized TM-specific methods in terms of SP and Total Column metrics. Moreover, qualitative topological analyses confirm that TMP-M2Align preserves the integrity of transmembrane helices and loop boundaries more effectively than competing approaches. These findings highlight the effectiveness of integrating topology-aware scoring with multiobjective optimization for achieving accurate and biologically meaningful alignments of TMPs.

Keywords: multiple sequence alignment; transmembrane proteins; multiobjective optimization; evolutionary algorithms; GPCR

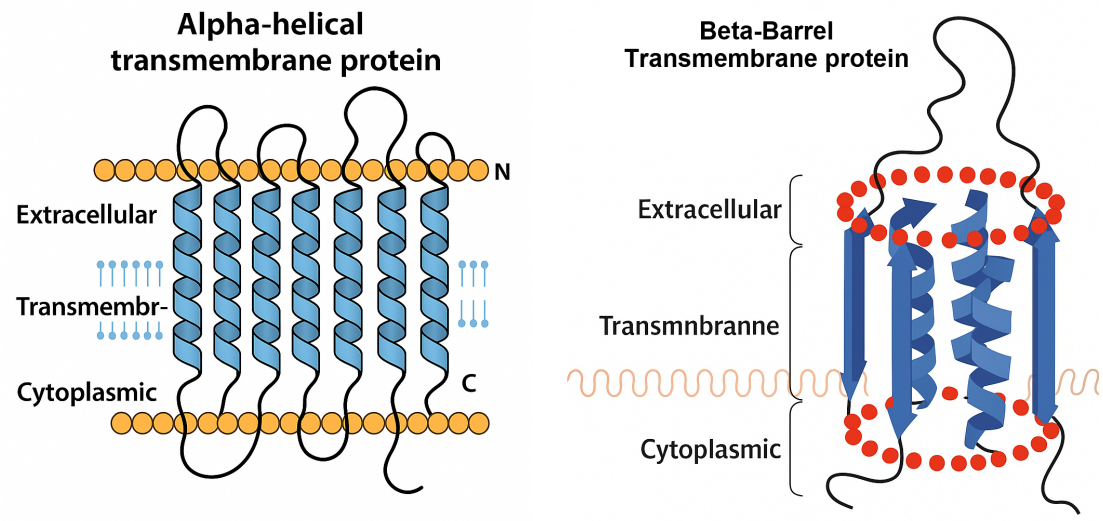
1. Introduction

Transmembrane proteins (TMPs) are nonsoluble proteins anchored in a cell membrane and containing one or more membrane-spanning segments separated with variable length intra or extracellular domains. Around 30% of the proteins encoded by the mammalian genome are transmembrane proteins [1,2]. Transmembrane proteins (TMPs) have attracted considerable attention playing a fundamental role in cell biology and are among the foremost tended to targets of pharmaceutical and life-science research. They are involved in a wide variety of vital biological processes, including transport of water-soluble molecules, flow of information, and energy production [3], and carry out fundamental capacities in numerous cellular and physiological processes, such as cell-cell recognition, molecular transport, and signal transduction. Given the biomedical importance of TMPs, sequence analysis methods are very significant.

Transmembrane proteins are divided into two main types: transmembrane alpha-helical proteins (αTMPs) and transmembrane beta-barrel proteins (βTMPs); this classification is based on the secondary

 doi:10.20944/preprints202510.0109.v1

structures that make up these proteins. The αTMPs are different from the soluble proteins in their conformation characteristics, as they have one or more helices that form helical bundles that cross the biological membrane. These transmembrane helixes (TMHs) are more hydrophobic than the helixes in soluble proteins. Each TMH corresponds to the membrane-spanning region of the protein sequence. The remaining parts of the sequence, referred to as non-TM regions, are classified as either inside (cytoplasmic) or outside (extracellular) segments, depending on their position relative to the lipid bilayer [4]. Figure 1 illustrates the structural differences between the two main types of transmembrane proteins. Figure 1a (left) shows an alpha-helical transmembrane protein (αTMP), characterized by multiple hydrophobic alpha-helices that span the lipid bilayer. These are common in eukaryotic organisms and are especially prevalent in receptor families such as G protein-coupled receptors (GPCRs). Figure 1b (right) presents a beta-barrel transmembrane protein (βTMP), typically found in the outer membranes of Gram-negative bacteria and certain organelles. The βTMPs form cylindrical structures composed of antiparallel beta-strands. Given these structural considerations, we focus in this paper exclusively on alpha-helical transmembrane proteins (αTMPs) due to their abundance in eukaryotic organisms and their critical biomedical roles.



(**a**) Alpha-helical transmembrane protein (αTMP). (**b**) Beta-barrel transmembrane protein (βTMP).

Figure 1. Comparison between the two main classes of transmembrane proteins: (a) αTMPs, composed of hydrophobic alpha-helices spanning the lipid bilayer; (b) βTMPs, cylindrical structures formed by antiparallel beta-strands.

As with soluble proteins and other biological sequences (DNA, RNA), alignment of two or more TMP sequences to identify regions of similarity, is an essential prerequisite for many modes of analysis in protein families such as homology modeling, secondary structure prediction, phylogenetic reconstruction, or the delineation of conserved and variable sites within a family [5]. This process is known as multiple sequence alignment (MSA), and it is a fundamental task in bioinformatics as the identification of similarity regions may indicate functional, structural, or evolutionary relationships between the sequences. The alignment of homologous sequences is crucial for understanding the evolutionary relationships among proteins and for predicting their functions based on sequence conservation.

Nowadays, many multiple sequence alignment (MSA) methods have been developed and extensively tested for aligning homologous soluble proteins. Notable examples include *ClustalW* [6], which uses progressive alignment with position-specific gap penalties and sequence weighting; *Kalign* [7], known for its high speed and accuracy based on Wu-Manber string-matching; *MAFFT* [8], which employs Fast Fourier Transform for rapid alignment; *MUSCLE* [9], which balances accuracy and scalability; and *T-Coffee* [10], which integrates multiple alignment sources through a consistency-based

 doi:10.20944/preprints202510.0109.v1

approach. While these tools are widely adopted, they are primarily optimized for globular proteins and often fail to capture the structural and topological constraints inherent to transmembrane proteins (TMPs). Although these methods can be applied to TMPs, they often yield significantly lower alignment accuracy compared to alignments of soluble proteins. This limitation arises because TMPs exhibit unique structural and topological features that are not adequately addressed by conventional MSA tools, necessitating the development of specialized approaches tailored to their characteristics [3].Despite the importance of TMPs, only a limited number of MSA methods have been specifically developed to address their unique structural characteristics. Notable examples include TM-Coffee [11], PRALINETM [5], and TM-Aligner [2]; however, these tools remain relatively few compared to those available for soluble proteins.

1.1. Related Work on TMP-Specific MSA Methods

Several methods have been specifically developed to align αTMPs, considering their unique structural and topological features. Early contributions include the approach of Cserzö *et al.* in [12], which used the positions of TMHs in one protein to identify corresponding helices in homologous sequences, providing a foundation for homology modeling of G protein-coupled receptors (GPCRs),such as human rhodopsin and bacteriorhodopsin [13]. The STAM (Simple Transmembrane Alignment Method) method [14] introduced substitution matrices and gap penalties tailored to TMPs, becoming the first dedicated software for their alignment. Forrest *et al.* [15] explored bipartite substitution schemes combining general and TMP-specific matrices, though with limited accuracy gains. Their study introduced HOMEP, a benchmark dataset of homologous membrane protein structures, which remains a valuable resource for assessing TMP alignment strategies [13].

Subsequent efforts incorporated homology information and topology prediction. TM-Coffee [11], an adaptation of T-Coffee, integrates homology extension with consistency-based alignment for αTMPs. PRALINE^{TM} [5] extends PRALINE by including topology predictions, the PHAT substitution matrix, and progressive consistency iterations. TM-Aligner [2] is a Web-based tool that uses dynamic programming to align TM and non-TM regions independently before merging them, while the TMSA approach [4] uses segment information extracted from topology predictions and evolutionary data to align TMPs. This sequential, segment-based approach emphasizes the structural and functional relevance of aligned regions.

Although these specialized methods improve alignment quality relative to general-purpose tools, their availability and performance remain limited. In particular, existing approaches often rely on fixed heuristics and lack flexibility to balance multiple alignment quality criteria simultaneously. This motivates the development of new methods that integrate topology-aware information into multiobjective optimization frameworks.

1.2. Contributions

In this study, we have adapted the multiobjective metaheuristic M2-Align [16] to the specific problem of aligning TMPs. Our implementation is based on the software framework introduced in TM-MSAligner [17], a tool specifically designed for multiple sequence alignment of TMPs. The proposed method simultaneously optimizes two widely used objective functions: (i) the Sum-of-Pairs (SOP) score with topology-aware gap penalties, and (ii) the TM-Regions Align (SA) score. Both scores were adapted to the characteristics of TMPs. In particular, the SOP function incorporates a region-aware gap penalty scheme and dynamically applies the PHAT substitution matrix within transmembrane segments. Higher penalties are imposed for opening or extending gaps in TM regions to preserve the structural integrity of conserved helices.

The main contributions of this work are as follows: 1. We present a novel multiobjective approach tailored for the alignment of transmembrane proteins. 2. Our method integrates topology-aware scoring mechanisms to enhance alignment accuracy. 3. We provide a comprehensive evaluation of our approach against existing methods using benchmark datasets. 4. We demonstrate the effectiveness

 doi:10.20944/preprints202510.0109.v1

of our method in aligning human G protein-coupled receptors (GPCRs) from classes A, B1, and C, showcasing its practical applicability in real-world scenarios.

The rest of the paper is organized as follows: Section 2 describes the proposed method, including topology prediction, scoring schemes, and the software framework. Section 3 presents the experimental setup, including datasets, evaluation metrics, and baseline methods. Section 4 discusses the results of our experiments, and Section 6 concludes the paper and outlines future work.

2. Materials and Methods

2.1. Overview of the Alignment Method

We propose a topology-aware multiobjective formulation for the multiple sequence alignment (MSA) of transmembrane proteins (TMPs). The method builds upon the TM-MSAligner framework [17], an open-source Java tool that implements a multiobjective evolutionary algorithm specifically adapted for TMPs. The alignment problem is modeled as a bi-objective optimization problem, where two quality scores are maximized simultaneously:

1. Sum-of-Pairs (SOP): This score is adapted to include topology-aware gap penalties and region-specific substitution matrices. It ensures that conserved regions, particularly transmembrane helices, are aligned accurately while penalizing gaps in these regions more heavily to preserve structural integrity.

2. Aligned Regions (AR): This score rewards topologically consistent alignments by prioritizing the alignment of residues within the same topological regions (e.g., transmembrane, intracellular, or extracellular). It ensures that functionally and structurally relevant domains are aligned cohesively.

By simultaneously optimizing these two objectives, the method generates a Pareto front approximation of alignment solutions. This allows researchers to select alignments that best suit their specific biological questions, balancing trade-offs between structural conservation and broader residue coverage.

2.2. Topology Prediction

Topology prediction is a critical step in adapting scoring strategies to the unique characteristics of TMPs. Accurate topology predictions enable the identification of transmembrane regions, intracellular loops, and extracellular loops, which are essential for designing topology-aware scoring schemes and gap penalties.

Several computational methods have been developed for topology prediction, with Hidden Markov Model (HMM)-based approaches such as TMHMM [18] and HMMTOP [19] achieving notable success. These methods leverage statistical models to predict the likelihood of residues belonging to specific topological regions. Additionally, machine learning-based methods have shown promising results, incorporating features such as sequence profiles and evolutionary information [20,21].

Recent advancements in deep learning have led to the development of more sophisticated models that can capture complex patterns in protein sequences. These models leverage large-scale sequence databases and transfer learning techniques to improve prediction accuracy. For example, DeepTMHMM [22] employs a deep learning architecture to predict transmembrane topology with high accuracy, outperforming traditional HMM-based methods. In this work, we utilize DeepTMHMM for topology prediction, as it has been shown to achieve state-of-the-art performance in predicting transmembrane regions and their orientations. DeepTMHMM uses a convolutional neural network (CNN) architecture to learn hierarchical features from protein sequences, enabling it to capture complex patterns associated with transmembrane topology. The model is trained on a large dataset of experimentally validated TMPs, allowing it to generalize well to novel sequences. DeepTMHMM provides per-residue annotations, including signal peptide (S), inside (I), alpha membrane (M), beta membrane (B), periplasm (P), and outside (O) regions. These annotations are directly integrated into the alignment pipeline and inform both substitution scoring and gap penalty decisions.

The use of DeepTMHMM offers several advantages:

 doi:10.20944/preprints202510.0109.v1

- High accuracy in predicting transmembrane topology, particularly for complex TMPs.
- Comprehensive per-residue annotations that inform scoring and gap penalties.
- Scalability to large datasets, making it suitable for high-throughput applications.

By incorporating topology predictions into our alignment framework, we ensure that the scoring and penalty schemes are tailored to the structural and functional constraints of TMPs. This integration significantly enhances the biological relevance and accuracy of the resulting alignments.

Although our framework depends on the accuracy of topology predictors, it is robust to moderate levels of misclassification. In practice, isolated errors in residue annotation are mitigated because substitution matrix assignment is applied at the column level, where consensus across multiple sequences reduces the weight of individual mistakes. Moreover, the Pareto-based optimization naturally prioritizes alignments that preserve both topological consistency and residue conservation, thereby down-weighting solutions that are negatively impacted by prediction noise. Importantly, DeepTMHMM has been shown to achieve particularly high accuracy in identifying transmembrane helices—the regions most critical for structural preservation—while errors are more likely to occur in loop regions, where alignment flexibility is biologically tolerable.

2.3. Topology-Aware Substitution Scoring

Traditional substitution matrices such as PAM [23] and BLOSUM [24] are widely used for MSA, but they are not ideal for transmembrane regions due to distinct evolutionary constraints [25]. The substitution score S_{ij} between amino acids i and j is computed based on background frequencies and observed substitution frequencies as:

$$S_{ij} = \frac{1}{\lambda} \ln \left(\frac{q_{ij}}{f_i f_j} \right) \tag{1}$$

To address the specific needs of TMPs, several specialized substitution matrices have been proposed, including JTT [26], PHAT [27], SLIM [28], and bbTM [29]. These matrices account for the unique physicochemical properties and evolutionary pressures of transmembrane regions. In our method, we dynamically assign substitution matrices based on the predicted topology of each residue:

- PHAT Matrix: Applied to residues within transmembrane (TM) regions, as it is specifically designed to capture the hydrophobic and structural constraints of TM helices.
- BLOSUM62 Matrix: Used for residues in non-transmembrane regions, such as intracellular and extracellular loops, to reflect the general evolutionary patterns of soluble proteins.

This dynamic assignment ensures that the scoring scheme is sensitive to the distinct characteristics of each topological region. Following the approach of PRALINE^{TM} [30] and TM-Aligner [2], the substitution matrix is selected for each column of the alignment based on the consistency of topology predictions across sequences. Columns with predominantly TM residues use the PHAT matrix, while columns with non-TM residues use BLOSUM62.

By tailoring substitution scoring to the topology of TMPs, our method improves the biological relevance and accuracy of alignments. This approach ensures that conserved transmembrane helices are aligned with high precision, while also accommodating the variability of loop regions.

2.4. Topology-Based Gap Penalty Scheme

A topology-aware gap penalty scheme is integrated into the SOP objective to preserve the structural integrity of conserved transmembrane regions. This scheme dynamically adjusts gap penalties based on the predicted topology of each residue, ensuring that gaps are penalized more heavily in transmembrane (TM) regions compared to non-TM regions.

The gap penalties are defined as follows:

- op_{tm} and ep_{tm}: Gap opening and extension penalties within TM regions.
- op_{non-tm} and ep_{non-tm}: Gap opening and extension penalties in non-TM regions.

 doi:10.20944/preprints202510.0109.v1

The dynamic gap penalty function $GapP(i)$ for residue i is computed as:

$$GapP(i) = \begin{cases} op_{tm} & \text{if } i = \text{GapOpen and } TP(i) = \text{TM} \\ op_{non-tm} & \text{if } i = \text{GapOpen and } TP(i) \neq \text{TM} \\ ep_{tm} & \text{if } i = \text{GapExtend and } TP(i) = \text{TM} \\ ep_{non-tm} & \text{if } i = \text{GapExtend and } TP(i) \neq \text{TM} \end{cases} \quad (2)$$

Here, $TP(i)$ represents the predicted topology region of residue i, which can be transmembrane (TM), intracellular, or extracellular. This approach ensures that gaps within TM helices are penalized more heavily, discouraging disruptions in these structurally critical regions. Conversely, gaps in non-TM regions, such as loops, are penalized less to allow for greater flexibility in aligning variable regions.

By incorporating topology-aware gap penalties, our method improves the biological relevance of alignments by preserving the continuity of TM helices and minimizing disruptions in conserved regions. This strategy is particularly effective for transmembrane proteins, where structural integrity is closely tied to functional performance.

2.5. Multi-Objective Formulation

The alignment of TMPs is formulated as a bi-optimization problem, where the goal is to find a set of alignments that optimize two complementary objectives: (i) the Sum-of-Pairs (SOP) score with topology-aware gap penalties and substitution matrices, and (ii) the Aligned Regions (AR) score, which rewards topologically consistent alignments. Since these objectives are conflicting, improving one generally entails a deterioration of the other. Consequently, the optimum is not a single solution but a set of trade-off, non-dominated solutions, known as the Pareto set, whose image in the objective space is the Pareto front. As the TMP-M2Align method is based on an evolutionary algorithm (see Section 2.6), its purpose is to approximate the Pareto front accurately. Each solution in this front corresponds to a possible alignment, thereby allowing biologists to select the option that best matches their specific requirements and preferences.

Both objectives are designed to be maximized and are specifically adapted to the unique structural and functional characteristics of TMPs. The alignment is represented as a set of k aligned sequences, $S = \{s_1, s_2, \ldots, s_k\}$, all of which have the same length L. The two objectives are defined in next sections.

2.5.1. Sum-of-Pairs with Gap Penalty and Topology prediction

The sum-of-pairs (SoP) score is a commonly used metric to evaluate the quality of a MSA. It measures the similarity between pairs of sequences in the MSA by summing the pairwise alignment scores of all possible pairs of sequences. In this study we have adapted the SOP score to consider the substitution matrix and gap penalties described in Section 2.3 and Section 2.4, respectively. The SOP score (Equation (3)) is computed by summing the pairwise scores of all sequences in the alignment, taking into account the predicted topology of each residue:

$$SOP(S) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} SP(s_i, s_j) \quad (3)$$

where $SP(s_i, s_j)$ represents the sum-of-pair score of the summing of the pairwise comparisons between each residue in each column of the s_i and s_j sequences as defined in Equation (4):

$$SP(s_i, s_j) = \sum_{l=1}^{L} \delta(s_{i,l}, s_{j,l}) - GP(s_{i,l}, s_{j,l}) \quad (4)$$

In Equation (4), δ represents the substitution matrix used: if both $s_{i,l}$ and $s_{j,l}$ residues in the column l are predicted to be member of a TM region, the PHAT matrix is used, otherwise BLOSUM62

 doi:10.20944/preprints202510.0109.v1

is used. This is done to guarantee that inconsistently predicted positions do not negatively influence the alignment quality.

The gap penalty value $GP(s_{i,l}, s_{j,l})$ is settled for determining the cost of aligning an amino acid with a gap, the gap penalty schema based on topology prediction is defined as follows:

$$GP(i,j) = \begin{cases} 0 & if\ i = j = \text{“}-\text{”}(gap) \\ GapP(i) & otherwise \end{cases} \tag{5}$$

where i and j represent the residues $s_{i,l}$, $s_{j,l}$ respectively. If both are not gaps, the gap penalty considering TM regions is calculated; this value is represented by $GapP(i)$ and defined in Equation (2). Otherwise, there is no penalty between i and j residues.

2.5.2. Aligned Regions

With the aim of preferentially aligning the regions with the same type, we have considered the second objective function: Aligned Regions, is computed by adding all the scores of the region compatibility between two sequence in each column of the alignment, is defined as:

The second objective function, Aligned Regions, is aimed at aligning preferentially regions with the same type. It is computed by adding all the scores of the region compatibility between two sequence in each column of the alignment, as defined in Equation (6):

$$AR(S) = \sum_{l=1}^{L} ARl(l) \tag{6}$$

where $ARl(l)$ is the Aligned Regions score of l column of the alignment, that is defined as follows:

$$ARl(l) = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} Match(s_{i,l}, s_{j,l}) \tag{7}$$

In Equation (7), $Match(s_{i,l}, s_{j,l})$ is the match score of two residues considering the type of topology and aminoacid, that is defined as follows:

$$Match(i,j) = \begin{cases} 4 & if\ TP(i) = TP(j) = TM \text{ and } i = j \\ 2 & if\ TP(i) = TP(j) = TM \text{ and } i \neq j \\ 2 & if\ TP(i) = TP(j) \neq TM \text{ and } i = j \\ 1 & if\ TP(i) = TP(j) \neq TM \text{ and } i \neq j \\ 0 & otherwise \end{cases} \tag{8}$$

where i and j represent the residues $s_{i,l}$ and $s_{j,l}$, respectively. If both are from different regions, the match value is zero. $TP(i)$ represents the topology region to which residue i belongs.

2.6. Implementation Details

The proposed TMP-M2Align method is implemented using TM-MSAligner [17], a flexible software framework developed for the multiple sequence alignment of transmembrane proteins using multiobjective evolutionary algorithms. TM-MSAligner is written in Java and leverages the jMetal framework [31,32], which provides a robust foundation for designing and executing customizable multiobjective metaheuristics.

The core of the optimization procedure is based on the NSGA-II multi-objective evolutionary algorithm [33] and consists of modular components for initialization, selection, crossover, mutation, and evaluation. Some components are customizable to accommodate different optimization strategies. Key components include:

- **Initialization**: The initial population is generated using a hybrid strategy inspired by the T-Coffee approach [34], where alignments produced by widely-used external tools such as ClustalW,

 doi:10.20944/preprints202510.0109.v1

Kalign, and MAFFT are merged and recombined to seed the population. This strategy provides a diverse and high-quality starting point for the evolutionary algorithm, helping it converge more efficiently towards accurate alignments.

- **Selection**: Supports random and tournament selection. In the tournament case, the tournament size can be tuned to control selection pressure.
- **Crossover**: Single Point Crossover (SPX) is used to combine parent alignments, promoting diversity in the offspring population.
- **Mutation**: Several biologically inspired mutation operators are available, such as InsertRandomGap (IRG), MergeAdjacentGapGroups (MAGG), ShiftClosedGaps (SCG), and SplitANonGapGroup (SANGG).
- **Evaluation**: TM-MSAligner allows running in both sequential and parallel execution modes, which is particularly useful for high-dimensional datasets.
- **Topology-aware evaluation**: The objective functions (Sum-of-Pairs and Aligned Regions) are integrated with transmembrane topology prediction. They dynamically apply region-specific substitution matrices and penalize gaps differently in TM and non-TM regions, as described in Sections 2.3 and 2.2.

Unlike structural homology-based methods such as TM-Aligner [2], TM-MSAligner does not rely on structural templates or 3D models. Instead, it uses only sequence data and predicted topology information, which reduces computational costs and broadens its applicability to large and uncharacterized datasets.

3. Experimental Setup and Benchmark Datasets

This section describes the datasets, evaluation metrics, baseline methods, and algorithmic configuration used to assess the performance of our multiobjective alignment approach.

3.1. Benchmark Datasets

We evaluated our method using Reference Set 7 of the BAliBASE 3.0 benchmark [35], which is a widely recognized standard for assessing the quality of multiple sequence alignment algorithms. This reference set is specifically designed for transmembrane proteins and includes eight families: `7tm`, `msl`, `dtd`, `acr`, `photo`, `ion`, `nat`, and `ptga`, comprising a total of 435 sequences. Each sequence contains between 2 and 14 transmembrane α-helices. The core domains are authors-defined, examining the alignment of structurally equivalent residues only. The main goal of BAliBASE is to assess the capacity of the strategies to recapitulate these core domains, mostly made of α-helices. Furthermore, it contains a program to assess the accuracy of candidate alignments over reference alignments provided by the benchmark, called Baliscore, that includes two metrics: Total Column (TC) and Sum-of-Pairs (SP) scores.

In addition to BAliBASE, we validated our approach on a real-world dataset composed of all human G protein-coupled receptors (GPCRs), one of the largest and most functionally diverse families of transmembrane proteins. The dataset includes representatives from three major GPCR classes: A (rhodopsin-like), B1 (secretin receptor family), and C (metabotropic glutamate/pheromone receptors). Protein sequences and topology annotations were obtained from the GPCRdb database [36]. This dataset allows us to test the scalability and biological relevance of our method on large, functionally important protein families.

3.2. Evaluation Metrics

We used two widely accepted metrics to evaluate the quality of the alignments:

- **Sum-of-Pairs (SP) Score:** This metric measures the proportion of residue pairs that are correctly aligned in the test alignment with respect to a reference alignment. Higher SP values indicate improved local homology conservation at the residue level, reflecting the algorithm's ability to align homologous residues consistently.

 doi:10.20944/preprints202510.0109.v1

- **Total Column (TC) Score:** This metric evaluates the number of alignment columns that are perfectly conserved with the reference, meaning that all residues in those columns are correctly aligned. High TC values indicate that the method preserves structurally and functionally coherent blocks across sequences, capturing entire regions rather than isolated pairs of residues.

Both SP and TC are particularly suitable for transmembrane proteins, as they simultaneously assess (i) the conservation of residues within hydrophobic transmembrane helices—critical for maintaining structural stability in the lipid bilayer—and (ii) the preservation of full topological domains such as helices and intra/extracellular loops, whose continuity is essential for protein function. These metrics were calculated using the `baliscore` tool provided by the BAliBASE package.

3.3. Compared Methods

To evaluate the performance of our approach, we conducted a comparative analysis against a set of widely used multiple sequence alignment (MSA) tools. These methods were selected because they represent a diverse range of alignment strategies, including traditional progressive aligners, consistency-based methods, and tools specifically adapted for transmembrane protein sequences. This selection allows us to assess both the general alignment performance and the effectiveness of topology-aware optimizations. The methods included in the comparison are summarized in Table 1.

Table 1. Summary of MSA tools used in the comparison. TMP = transmembrane protein.

Tool	Type	TMP Support
ClustalW [6]	Progressive	No
MAFFT [8]	Progressive	No
Kalign [7]	Progressive	No
Muscle [9]	Progressive	No
T-Coffee [10]	Consistency-based	No
TM-Coffee [11]	Consistency-based	Yes
PRALINETM [5]	Consistency-based	Yes
TM-Aligner [2]	Heuristic	Yes

All baseline tools were executed using their default parameters, unless specific options were recommended in their documentation for handling transmembrane protein sequences. TM-Coffee and PRALINETM were included for their explicit consideration of transmembrane topology. TM-Aligner was originally intended as a comparison method, but it has not been included due to the continued inaccessibility of its Web server during the evaluation period.

3.4. TMP-M2Align Configuration

Our TMP-M2Align implementation was configured based on the parameters shown in Table 2, following the executable setup used in our experiments. The algorithm was executed with a limit of 25,000 evaluations and a population size of 100 individuals.

The predicted topologies were obtained using DeepTMHMM and stored in `.3line` files, which were parsed during runtime to dynamically inform the scoring process. Each benchmark folder contained a minimum of two precomputed alignments (in FASTA format), used to seed the initial population. These alignments were validated prior to execution.

In our configuration, the choice of gap penalty parameters was guided by the structural properties of transmembrane proteins. Transmembrane (TM) helices are highly conserved segments whose continuity is essential for maintaining the stability of the protein within the lipid bilayer. Therefore, we assigned stronger penalties for gap opening (8) and extension (3) in TM regions to discourage insertions or deletions that could disrupt these helices. In contrast, non-TM regions such as cytoplasmic and extracellular loops are more flexible and variable across homologs. For these regions, lower penalties for gap opening (3) and extension (1) were applied, allowing the aligner to introduce gaps where structural variability is expected. This topology-aware scheme balances the need to preserve conserved

 doi:10.20944/preprints202510.0109.v1

TM segments while accommodating sequence divergence in loop regions, and is consistent with previous studies that emphasize the differential treatment of TM and non-TM domains in alignment strategies.

Table 2. Configuration parameters used in TMP-M2Align during the experiments.

Parameter	Value
Max evaluations	25,000
Population size	100
Crossover probability	0.8
Mutation probability	0.2
Mutation operator	`ShiftClosedGapsMSAMutation`
Gap opening (TM regions)	8
Gap extension (TM regions)	3
Gap opening (non-TM regions)	3
Gap extension (non-TM regions)	1
Substitution matrix (TM)	PHAT
Substitution matrix (non-TM)	BLOSUM62
Objective functions	Sum-of-Pairs with Topology, Aligned Segments
Topology predictor	DeepTMHMM
Initial population source	ClustalW, Kalign, MAFFT alignments

All experiments, including those involving baseline alignment tools (ClustalW, Kalign, MAFFT, MUSCLE, T-Coffee, PRALINETM, and TM-Coffee), were conducted on a machine with an Intel Xeon CPU E5-2650 v3 @ 2.30GHz and 32 GB of RAM, running Ubuntu 20.04 LTS. This uniform setup ensures that performance comparisons across different methods were conducted under equivalent hardware and software conditions.

4. Results

4.1. BAliBASE Results

It is worth noting that, since TMP-M2Align produces a set of non-dominated solutions (i.e., a Pareto front), we selected for each BAliBASE instance the alignment with the highest SP score and the alignment with the highest TC score, and then reported their respective scores in Table 3. This strategy ensures that the reported results reflect the best-case performance of TMP-M2Align with respect to each metric. The TM-specific methods TM-Coffee and TM-Aligner are included as references, although they were not rerun due to technical limitations (see Section 3.3). Regarding the use of traditional tools such as MAFFT, ClustalW, and Kalign—which are not specifically designed for transmembrane proteins—it is important to clarify that all of them were able to generate syntactically valid alignments. These alignments were evaluated using BAliBASE's `baliscore` tool, which assesses structural alignment quality based on reference core domains. However, their lack of topology awareness inherently limits their performance on TMP datasets, as reflected in the lower TC and SP scores compared to methods designed or adapted for transmembrane sequences.

Table 3. SP and TC scores for each instance in the BAliBASE Reference Set 7. TM-Aligner and TM-Coffee scores were taken from reference results due to server inaccessibility.

Tool	ClustalW		Kalign		MAFFT		Muscle		Praline		T-Coffee		TM-Coffee		TM-Aligner	TMP-M2Align	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC	SP	SP	TC
msl	0.862	0.620	0.845	0.710	0.817	0.580	0.835	0.580	0.804	0.540	0.840	0.600	0.828	0.600	**0.880**	0.855	**0.720**
7tm	0.844	0.330	0.782	0.260	0.725	0.160	0.807	0.270	0.755	0.390	0.852	0.400	0.862	**0.420**	0.820	**0.864**	0.410
ion	0.462	0.060	0.469	0.020	0.444	0.000	0.480	0.000	0.417	0.000	0.511	0.050	0.506	0.050	0.510	**0.514**	**0.060**
photo	**0.947**	0.670	0.874	0.400	0.835	0.380	0.897	0.490	0.823	0.410	0.918	0.570	0.928	0.590	0.920	0.943	**0.670**
ptga	0.654	0.060	0.688	0.080	0.629	0.050	0.651	0.050	0.641	**0.190**	0.687	0.150	0.705	0.160	0.700	**0.718**	0.110
acr	0.910	0.650	0.880	0.470	0.901	0.490	0.942	0.670	0.876	0.470	0.929	0.680	0.930	0.660	0.920	**0.944**	**0.710**
nat	0.722	0.180	0.697	0.160	0.672	0.060	0.723	0.160	0.606	0.000	0.695	0.110	0.700	0.100	0.750	**0.755**	**0.220**
dtd	0.844	0.270	0.846	0.360	0.812	0.180	0.841	0.310	0.760	0.210	0.863	0.290	0.877	0.370	0.870	**0.879**	0.330
Average	0.781	0.355	0.760	0.308	0.729	0.238	0.772	0.316	0.710	0.276	0.787	0.356	0.792	0.369	0.796	**0.809**	**0.404**

 doi:10.20944/preprints202510.0109.v1

Our proposed method, achieved the highest average SP (0.809) and TC (0.404) scores, outperforming all compared tools. Notably, even the best-performing classical tools (T-Coffee and ClustalW) fell short in TC score, highlighting the benefit of using a topology-aware scoring strategy for transmembrane proteins. The improvement over TM-Coffee and TM-Aligner is also evident in both metrics.

Our results suggest that our proposal has more accuracy similar to the tools based on homology-modeling and TM-Aligner, however. These methods require template structures for alignment accuracy. However, our approach does not depend on templates.

4.2. GPCR Dataset Results

The results on the GPCR dataset, summarized in Table 4, further validate the effectiveness of our method. For each GPCR class (A, B1, and C), we compared TMP-M2Align with ClustalW, Kalign, MAFFT, Muscle, and TM-Aligner. Due to the lack of reference alignments for this dataset, we evaluated the alignments using the same SP and TC metrics employed in the BAliBASE benchmark. As with the previous evaluation, from the set of non-dominated solutions generated by TMP-M2Align, we selected the alignment with the highest SP score and the one with the highest TC score for each GPCR class. These selected alignments represent the best-performing solutions with respect to each quality criterion, and their scores are reported in the corresponding columns of Table 4.

Table 4. Sum-of-Pairs (SP) and Total Column (TC) scores for the best alignments obtained for each GPCR class. The results for TMP-M2Align correspond to the best-scoring solutions selected from the generated Pareto fronts. TMP-M2Align consistently outperforms traditional aligners across all three classes.

GPCR	ClustalW		Kalign		MAFFT		Muscle		TMP-M2Align	
	SP	TC	SP	TC	SP	TC	SP	TC	SP	TC
Class A	0.653	0.000	0.578	0.000	0.576	0.000	0.626	0.000	**0.663**	**0.020**
Class B1	**0.612**	0.550	0.589	0.490	0.586	0.530	0.582	0.530	0.611	**0.560**
Class C	0.652	0.190	0.519	0.000	0.677	0.360	0.506	0.000	**0.692**	**0.360**
Average	0.639	0.247	0.562	0.163	0.613	0.297	0.571	0.177	**0.655**	**0.313**

In GPCR Classes A and C, TMP-M2Align produced alignments that outperformed all traditional aligners in at least one of the evaluated metrics. Notably, in Class B1, it achieved the highest Total Column (TC) score while maintaining a competitive Sum-of-Pairs (SP) value, indicating improved residue conservation without sacrificing overall alignment quality. In Class C, TMP-M2Align outscored all other methods in both SP and TC, demonstrating its ability to generate structurally consistent and functionally relevant alignments. These results underscore the advantage of using multiobjective optimization to explore a diverse range of solutions, allowing practitioners to select alignments tailored to specific biological goals.

Due to technical limitations, some transmembrane-specific alignment tools could not be included in the GPCR dataset evaluation. In particular, the Web server for **TM-Aligner** (http://lms.snu.edu.in) was persistently unreachable, returning a *"connection timed out"* error. Similarly, attempts to run **TM-Coffee** and **PRALINE^{TM}** via their respective Web platforms resulted in *403 Forbidden* or access-denied responses. Since these tools are only available as Web services and do not provide standalone versions, it was not possible to include them in the GPCR experiments under the same computational conditions. Therefore, while their published results were considered for the BAliBASE benchmark, they were excluded from the GPCR dataset evaluation.

4.3. Visualization of Solution Fronts

TMP-M2Align does not return a single alignment but a set of non-dominated solutions distributed along a Pareto front, thus providing practitioners with the flexibility to select alignments that emphasize either highly conserved columns (Total Column, TC score) or broader homologous residue coverage (Sum-of-Pairs, SP score). Figures 2 and 3 show typical Pareto front approximations produced by TMP-M2Align for the GPCR Class A and Class B1 datasets, respectively. In these fronts, the extreme

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 October 2025 doi:10.20944/preprints202510.0109.v1

points correspond to solutions that maximize one objective at the expense of the other, while the remaining points represent trade-offs with different balances between the two objectives.

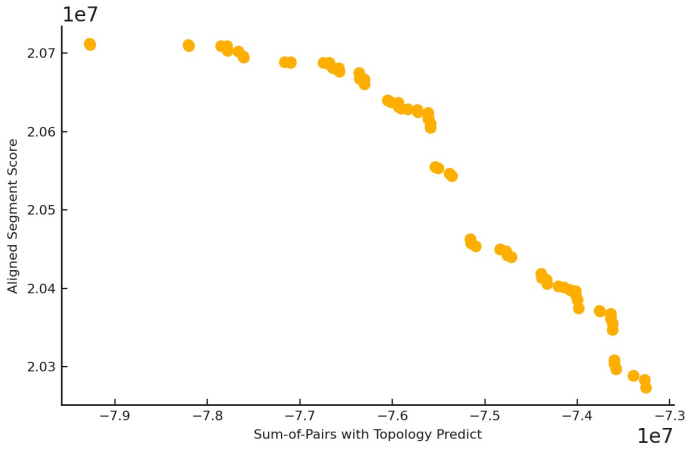


Figure 2. Pareto front approximation obtained by TMP-M2Align on the GPCR Class A dataset. Each point corresponds to a trade-off solution between SP and TC objectives.

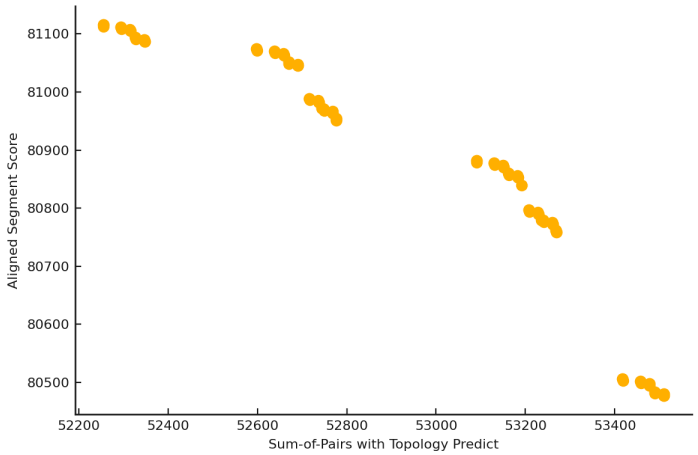
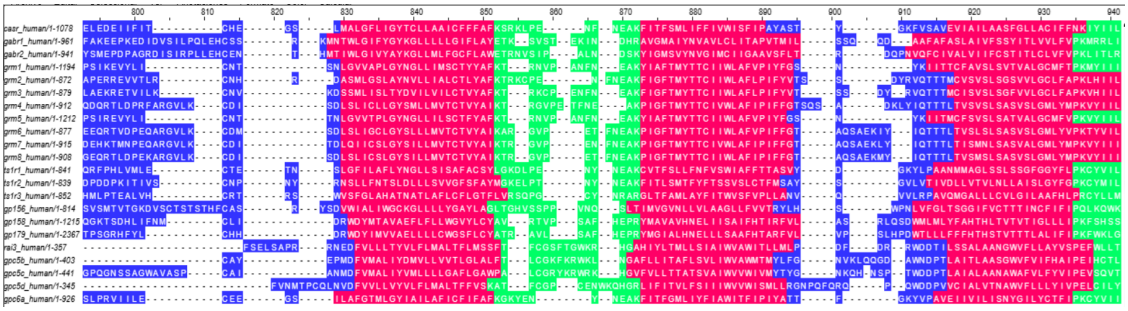


Figure 3. Pareto front approximation obtained by TMP-M2Align on the GPCR Class B1 dataset. The distribution reflects multiple high-quality alignments with varying emphasis on residue conservation versus topological consistency.

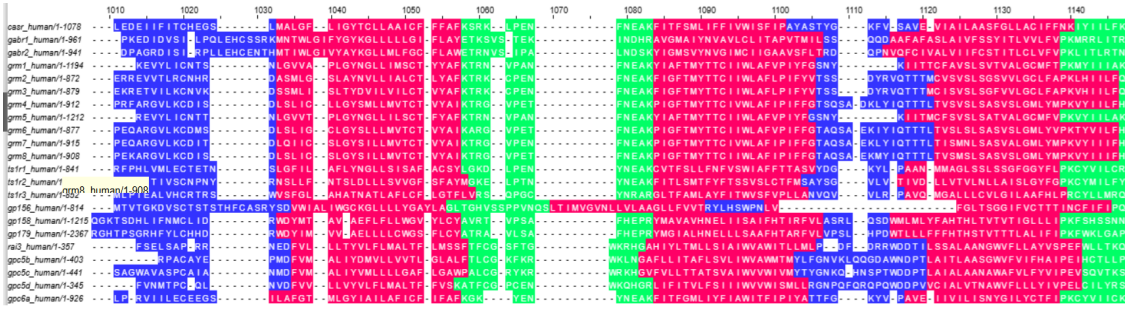
These fronts confirm that multiobjective optimization is effective at capturing the inherent trade-offs in MSA quality metrics. In GPCR Class A, the front is broad and well-structured, highlighting the algorithm's ability to explore diverse alignment solutions. For GPCR Class B1, the front appears tighter but still exhibits a clear gradient of trade-offs, suggesting the presence of structurally variable or functionally divergent regions that benefit from adjustable alignment strategies. This capability is particularly valuable for datasets where balancing structure and homology is critical.

To illustrate the practical meaning of the Pareto fronts, we selected three representative solutions from the GPCR Class C dataset: (i) the solution with the highest Aligned Regions score (BestARscore), (ii) the solution with the highest Sum-of-Pairs score (BestSOPscore), and (iii) an intermediate solution located in the middle of the front (Medium). These alignments are shown in Figure 4.

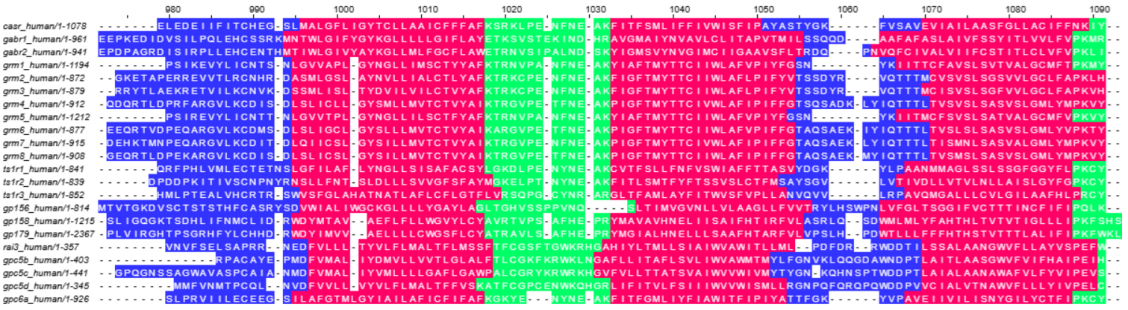
 doi:10.20944/preprints202510.0109.v1



(**a**) BestARscore solution



(**b**) BestSOPscore solution



(**c**) Medium solution

Figure 4. Representative solutions selected from the Pareto front of GPCR Class C. Residues are color-coded according to topology: red = transmembrane helices, green = intracellular regions, blue = extracellular regions. (a) BestARscore solution, (b) BestSOPscore solution, (c) Medium solution.

The BestARscore solution (Figure 4a) maximizes topological consistency: transmembrane helices appear as uninterrupted blocks and loop boundaries are preserved, although this alignment sacrifices some residue-level conservation across sequences. The BestSOPscore solution (Figure 4b) maximizes residue conservation, aligning a greater number of identical and similar residues. However, this comes at the cost of introducing gaps inside helices, fragmenting their continuity and reducing biological interpretability. Finally, the Medium solution (Figure 4c) provides a balanced trade-off, maintaining both a reasonable level of residue conservation and structural coherence across helices and loops.

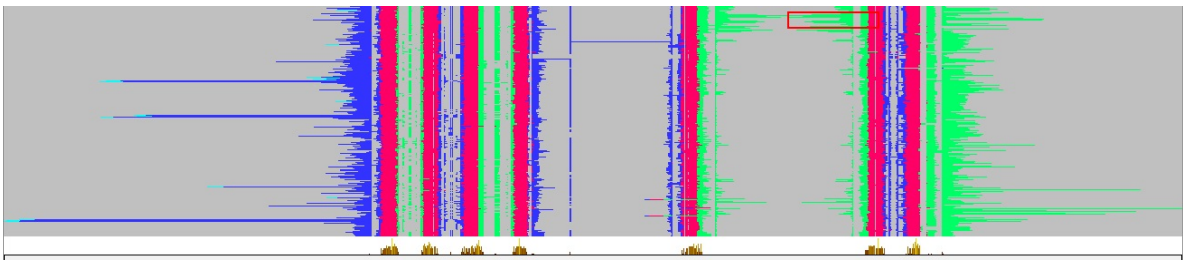
These examples demonstrate how the Pareto front captures distinct alignment strategies. Practitioners can therefore choose the solution that best fits their objectives: AR-oriented alignments for structural studies , SOP-oriented alignments for evolutionary or motif analyses, and medium trade-offs when both perspectives are relevant.

4.4. Topological Analysis of the Alignments in Class A and Class C GPCRs

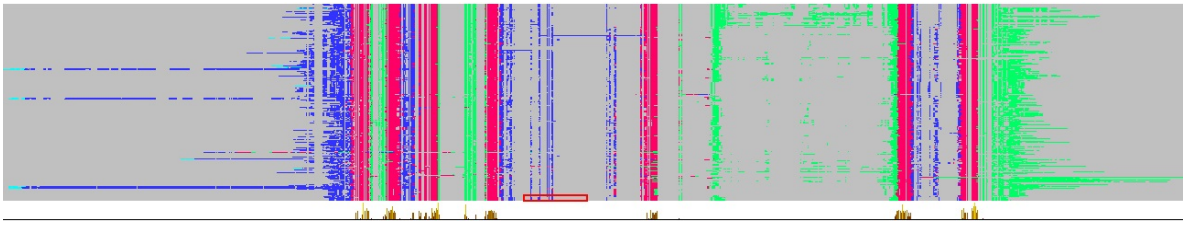
Figures 5 and 6 present a visual comparison of the multiple sequence alignments generated for Class A (Rhodopsin-like) and Class C (Glutamate-like) human GPCRs using various alignment tools. Per-residue topological annotations were predicted using *DeepTMHMM* and visualized in *Jalview*.

 doi:10.20944/preprints202510.0109.v1

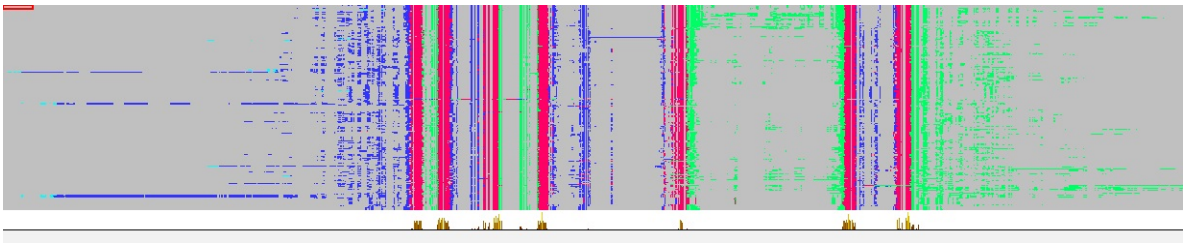
Color coding is applied to highlight structural regions: transmembrane helices (TM helix) in red, intracellular regions (inside) in green, and extracellular regions (outside) in blue.



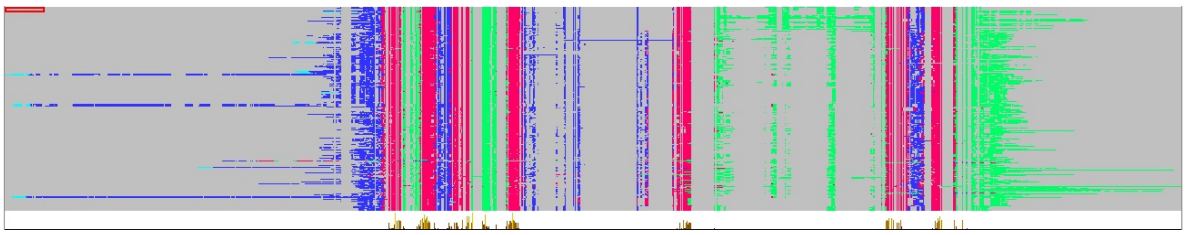
(**a**) Reference alignment



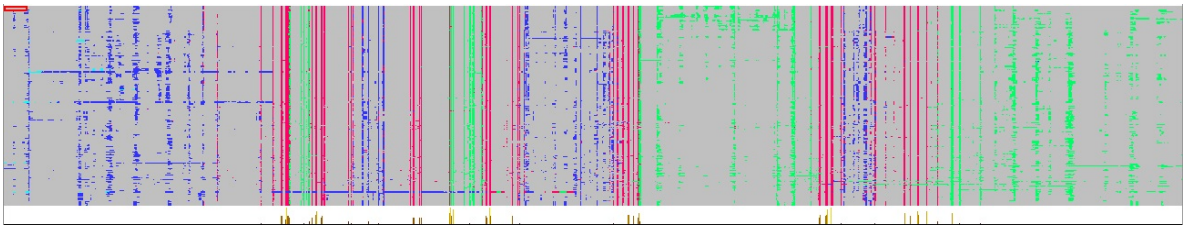
(**b**) TMP-M2Align



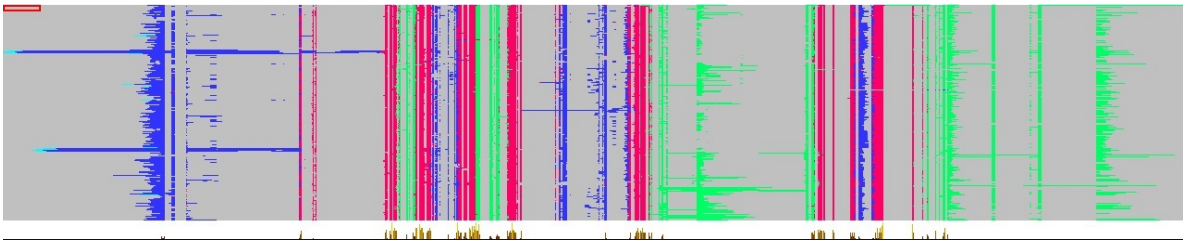
(**c**) ClustalW



(**d**) Kalign



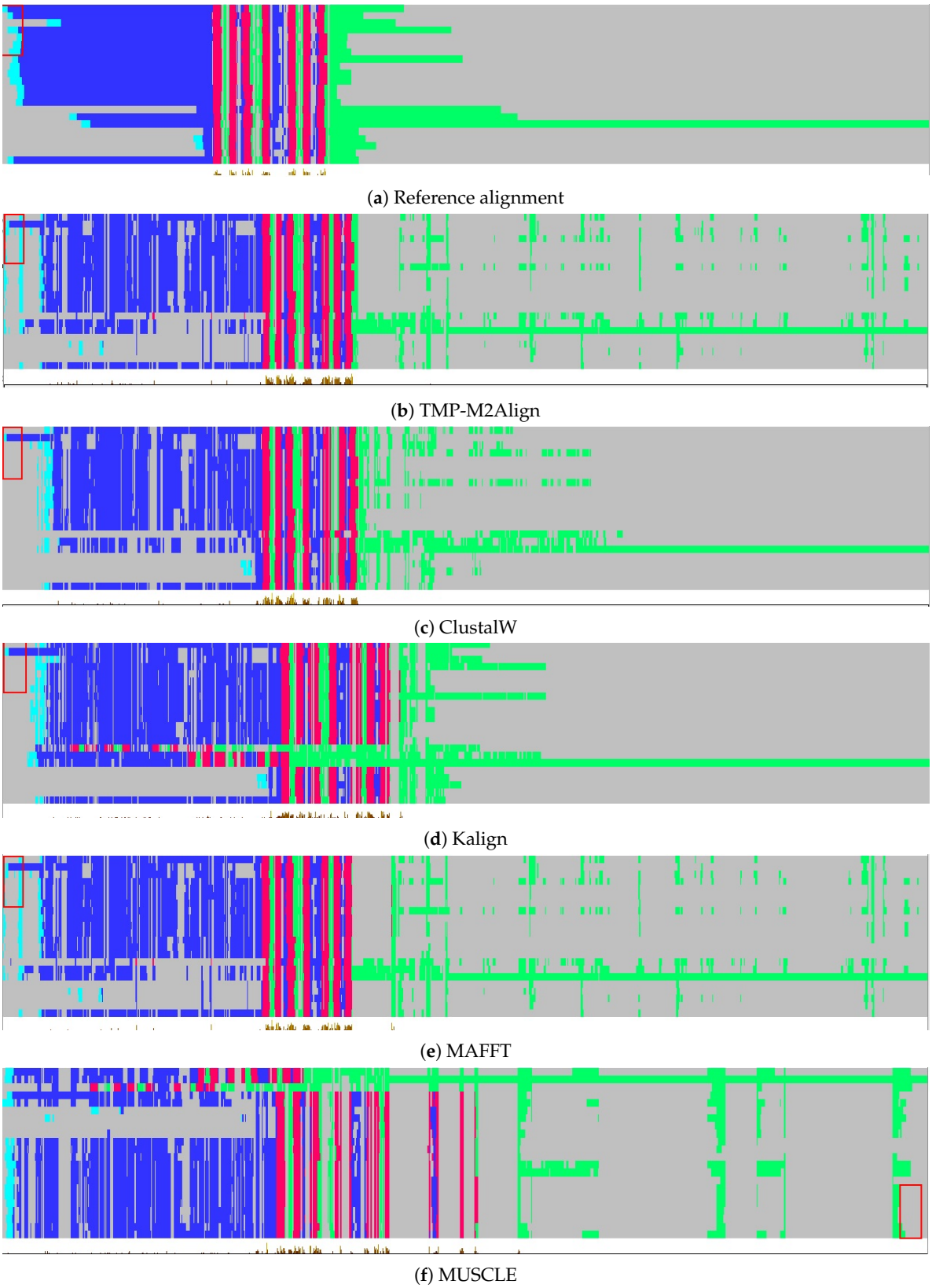
(**e**) MAFFT



(**f**) MUSCLE

Figure 5. Visual comparison of multiple sequence alignments for Class A (Rhodopsin-like) GPCRs. Topological annotations predicted by *DeepTMHMM* are visualized in *Jalview* using the color code: red (TM helix), green (inside), and blue (outside).

 doi:10.20944/preprints202510.0109.v1



(**a**) Reference alignment

(**b**) TMP-M2Align

(**c**) ClustalW

(**d**) Kalign

(**e**) MAFFT

(**f**) MUSCLE

Figure 6. Visual comparison of multiple sequence alignments for Class C (Glutamate-like) GPCRs. Topological annotations predicted by *DeepTMHMM* are visualized in *Jalview* using the color code: red (TM helix), green (inside), and blue (outside).

In both classes, the reference alignment, obtained from *GPCRdb*, exhibits a clearly defined transmembrane architecture with compact and uninterrupted TM segments, flanked appropriately by inside and outside loops. These topologies serve as reliable structural ground truths for assessing

 doi:10.20944/preprints202510.0109.v1

the biological validity of each alignment. Our method, TMP-M2Align, achieves high topological consistency in both classes, producing alignments that preserve domain continuity and minimize unwanted gaps within TM segments. The benefits of our approach are particularly evident in Class C, which is more challenging due to higher evolutionary divergence and longer sequence lengths. TMP-M2Align aligns the transmembrane regions as continuous blocks and maintains the separation of topological domains better than the other methods.

By contrast, traditional alignment tools such as ClustalW, Kalign, MAFFT, and MUSCLE exhibit greater fragmentation and inconsistency in both classes. In Class A, some TM helices are reasonably preserved, but gaps and loop misalignments are common, especially with MAFFT and MUSCLE. In Class C, the issues are more severe: transmembrane helices are frequently interrupted by internal gaps, and inside/outside regions are scattered throughout the alignment.

Overall, TMP-M2Align demonstrates superior robustness across both GPCR classes. It consistently preserves the structural integrity of topological segments, making it more suitable for downstream structural and functional analyses involving transmembrane proteins.

4.5. Representative Alignment Snippets

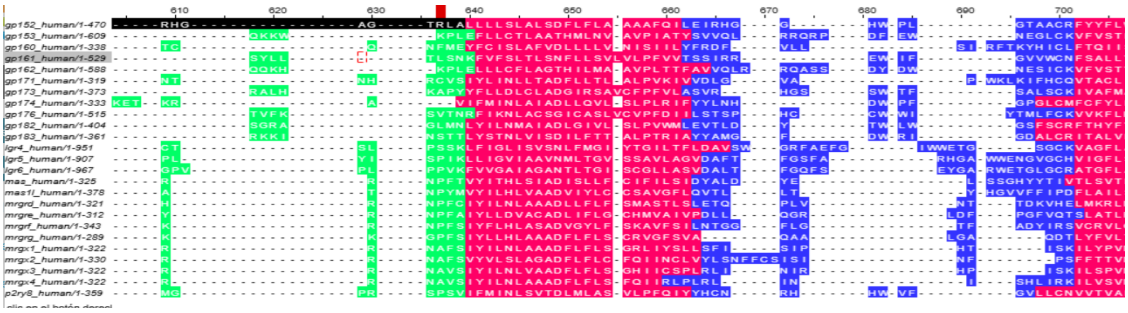
To further illustrate the impact of our topology-aware scoring and penalty scheme, we show in Figure 7 an excerpt of a representative transmembrane segment and its flanking loop regions. This region was selected from the GPCR Class A dataset, as it provides a clear example of the structural constraints typical of transmembrane helices. The reference alignment obtained from GPCRdb is presented in Figure 7a, while the alignment produced by our TMP-M2Align algorithm is shown in Figure 7b. In both cases, residues are colored according to their predicted topology: transmembrane helices in red, intracellular regions in green, and extracellular regions in blue. Figure 7c, Figure 7d and Figure 7e present the results obtained with MAFFT, ClustalW and Kalign, respectively.

As shown in Figure 7a, the reference alignment preserves the general topology of the region but contains a significant number of internal gaps, particularly within transmembrane helices, which fragment these highly conserved domains. By contrast, TMP-M2Align produces a more compact alignment in which gaps are largely confined to loop regions, while the integrity of the helical transmembrane segments is maintained Figure 7b. This behavior is a direct consequence of our topology-aware gap penalty scheme, which imposes stronger penalties for insertions or deletions within transmembrane regions. Additionally, the *Aligned Regions* objective function contributes to the coherent placement of loop residues, reducing misalignments across cytoplasmic and extracellular domains.

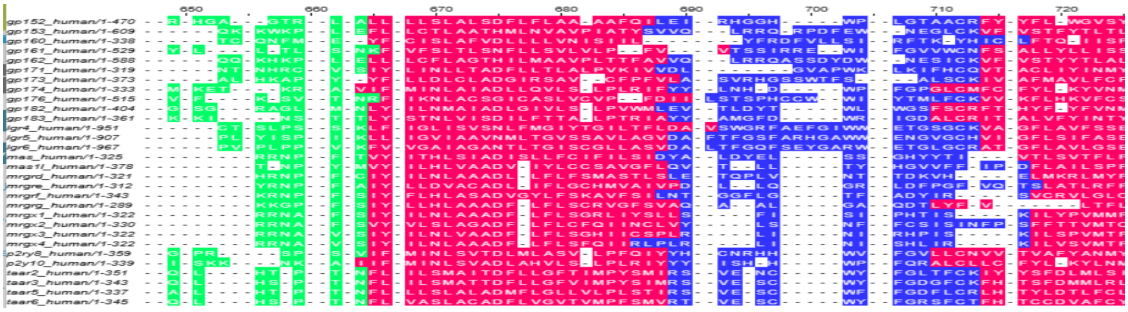
The comparison with other alignment tools highlights the advantages of our approach. MAFFT tends to lose alignment consistency at loop boundaries, fragmenting conserved helices (Figure 7c). ClustalW partially preserves helices but scatters gaps across both loops and transmembrane regions due to its lack of topology awareness (Figure 7d). Kalign maintains the separation of topological domains but does not effectively reduce the number of gaps within helices, leading to unnecessary fragmentation (Figure 7e).

Overall, TMP-M2Align not only improves quantitative metrics such as SP and TC scores but also yields biologically meaningful alignments where helices remain continuous and loop boundaries are well defined. This reinforces the value of explicitly integrating topology into the alignment process, enabling our method to deliver alignments that are structurally faithful and suitable for downstream applications such as homology modeling and functional annotation.

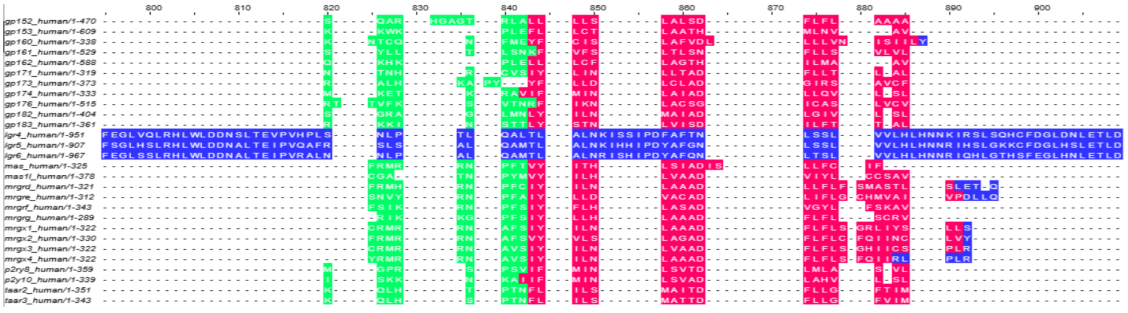
 doi:10.20944/preprints202510.0109.v1



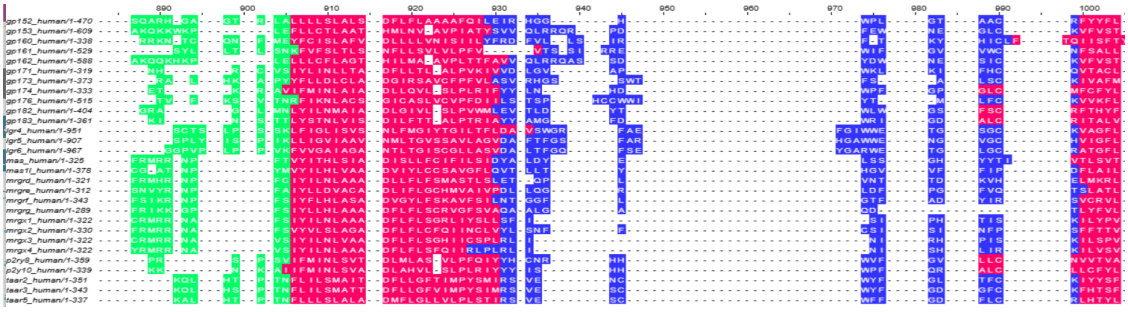
(**a**) Reference alignment (GPCRdb)



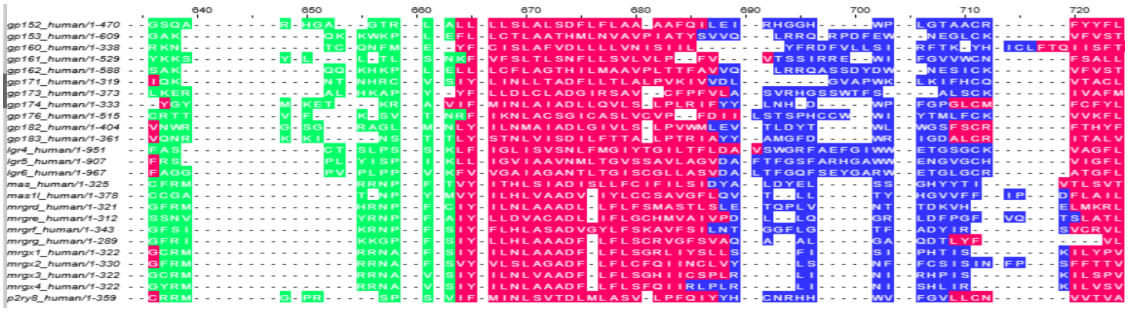
(**b**) TMP-M2Align



(**c**) MAFFT



(**d**) ClustalW



(**e**) Kalign

Figure 7. Representative alignment snippets from GPCR Class A. Color scheme: transmembrane helices (red), intracellular regions (green), and extracellular regions (blue).

 doi:10.20944/preprints202510.0109.v1

5. Discussion

The results obtained in this study demonstrate that our approach is an effective strategy for the multiple sequence alignment of transmembrane proteins, especially for highly divergent families such as GPCRs. Its performance surpasses that of classical alignment tools and even specialized TM tools in both BAliBASE and real-world GPCR datasets.

5.1. Implications of a Multiobjective Framework for TMPs

The use of a multiobjective optimization strategy allows simultaneously maximize multiple alignment quality metrics. This is particularly beneficial for TMPs, where trade-offs often exist between structural consistency and alignment completeness. By generating an approximated Pareto front of solutions, our approach empowers users to select the alignment most appropriate for their biological question, whether they prioritize conservation, coverage, or a balance between both.

Consequently, the superior SP and TC values obtained by TMP-M2Align confirm that the multi-objective approach not only enhances residue-level conservation but also preserves the topological integrity of transmembrane domains, making the alignments more biologically meaningful.

5.2. Advantages of Topology-Aware Gap Penalties

One of the core innovations in our approach is its use of topology-aware scoring and gap penalty schemes. The algorithm increases gap penalties within predicted TM regions, thus discouraging disruptive insertions or deletions that would break the continuity of membrane-spanning helices. This simple yet powerful modification substantially improves alignment quality by preserving biologically meaningful regions.

Additionally, different substitution matrices are dynamically applied depending on the predicted topological region, using PHAT in TM helices and BLOSUM62 elsewhere. This contributes to better biological relevance by reflecting region-specific evolutionary pressures.

5.3. Comparison with Existing Methods

The superiority of our approach can be attributed to three main factors: (i) the integration of topological information into both scoring and gap penalty schemes, (ii) the dynamic use of substitution matrices depending on transmembrane predictions, and (iii) the use of metaheuristics, which explore diverse alignment hypotheses rather than relying on deterministic heuristics.

While tools like TM-Coffee and PRALINE^{TM} do incorporate topology information, their methods are fixed and lack flexibility. In contrast, TMP-M2Align dynamically adapts to predicted TM regions and uses evolutionary search to refine alignments based on competing criteria. This flexibility makes it more adaptable to different sequence families and alignment goals.

Consequently, the superior SP and TC values obtained by TMP-M2Align confirm that the multi-objective approach not only enhances residue-level conservation but also preserves the topological integrity of transmembrane domains, making the alignments more biologically meaningful.

5.4. Qualitative Topological Consistency of GPCR Alignments

Figures 5 and 6 allow a qualitative evaluation of alignment quality beyond quantitative metrics. For Class A GPCRs, TMP-M2Align produces alignments where transmembrane helices (blue) appear as continuous, uninterrupted segments, with very few artificial gaps introduced inside these regions. This demonstrates that our method preserves the structural integrity of the seven canonical helices while maintaining the clear separation of intracellular (red) and extracellular (green) loops. In contrast, baseline methods such as MAFFT and MUSCLE introduce fragmentation within helices and scatter loop residues across non-homologous positions, reducing biological interpretability.

In Class C GPCRs, which are characterized by greater sequence divergence and longer extracellular domains, the qualitative differences become more evident. TMP-M2Align maintains the block-like organization of transmembrane helices and respects the orientation of topological domains, whereas ClustalW and Kalign frequently disrupt helix continuity with insertions and misplace loop regions

 doi:10.20944/preprints202510.0109.v1

relative to the reference alignment. These distortions compromise the accurate delineation of structural domains and could negatively impact downstream structural or evolutionary analyses.

Taken together, the visual inspection of Figures 5 and 6 confirms that TMP-M2Align not only improves numerical scores such as SP and TC but also yields biologically meaningful alignments that preserve the continuity and organization of topological regions. This qualitative robustness is particularly valuable for GPCRs, where maintaining the integrity of helices and loop boundaries is critical for structure-function studies.

6. Conclusions and Future Work

This work aims to evaluate the effectiveness of a topology-aware, multiobjective approach for the multiple sequence alignment (MSA) of transmembrane proteins (TMPs), using both benchmark data (BAliBASE Reference Set 7) and a real-world dataset consisting of human GPCRs from different classes. To this end, we introduced TMP-M2Align, a software tool which integrates topological information into its scoring scheme and gap penalties.

Our results showed that TMP-M2Align consistently outperforms traditional alignment tools such as ClustalW, Kalign, MAFFT, and MUSCLE, as well as specialized TM tools like TM-Coffee and TM-Aligner (when available). It achieves higher alignment quality, as measured by Sum-of-Pairs (SP) and Total Column (TC) scores, particularly in challenging cases like GPCR Classes B1 and C. The use of a multiobjective optimization strategy enables the generation of Pareto front approximations, offering a spectrum of high-quality alignments that allow users to prioritize either topological consistency or broader residue matching.

One of the key innovations lies in the topology-aware scoring, which includes region-sensitive gap penalties and dynamic substitution matrices. These features contribute to preserving the integrity of transmembrane helices and topological domains, as demonstrated by qualitative assessments using DeepTMHMM annotations and visual inspection in Jalview.

Our approach relies on accurate topology predictions as a prerequisite, which introduces some dependency on external tools such as DeepTMHMM. In future work, the integration of confidence levels from topology predictors or structural validation from tools like AlphaFold2 could enhance robustness. Additionally, incorporating new objective functions based on structural similarity or co-evolutionary information may further expand its applicability.

In conclusion, TMP-M2Align offers a flexible, effective, and biologically relevant solution for aligning transmembrane proteins. Its ability to balance competing alignment goals makes it especially suitable for complex datasets like GPCRs, and its design opens pathways for further integration with structural bioinformatics tools and workflows.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: This work has been partially funded by grant PID2024-155363OB-C41 funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, KOSMOS (Boosting Federated Data-Sharing Ecosystems with a Multidimensional Approach) and by grant DGP_PIDI_2024_01174, Junta de Andalucia.

Acknowledgments: The authors would like to express their gratitude to the State Technical University of Quevedo for the support provided throughout this research. Their continuous encouragement and resources have been invaluable in the development of this study.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202510.0109.v1

References

1. Wallin, E.; Heijne, G.V. Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Science* **1998**, *7*, 1029–1038.
2. Bhat, B.; Ganai, N.A.; Andrabi, S.M.; Shah, R.A.; Singh, A. TM-Aligner: Multiple sequence alignment tool for transmembrane proteins with reduced time and improved accuracy. *Scientific Reports* **2017**, *7*. https://doi.org/10.1038/s41598-017-13083-y.
3. Frishman, D. *Structural Bioinformatics of Membrane Proteins*; Springer, 2010; p. 400. https://doi.org/10.1007/978-3-7091-0045-5.
4. Wang, H.; Wang, J.; Zhang, L.; Sun, P.; Du, N.; Li, Y. A sequential segment based alpha-helical transmembrane protein alignment method. *International Journal of Biological Sciences* **2018**, *14*, 901–906. https://doi.org/10.7150/ijbs.24327.
5. Pirovano, W.; Feenstra, K.A.; Heringa, J. PRALINE™: A strategy for improved multiple alignment of transmembrane proteins. *Bioinformatics* **2008**, *24*, 492–497. https://doi.org/10.1093/bioinformatics/btm636.
6. Thompson, J.; Higgins, D.; Gibson, T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **1994**, *22*, 4673–4680.
7. Lassmann, T.; Sonnhammer, E. Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* **2005**, *6*, 298. https://doi.org/10.1186/1471-2105-6-298.
8. Katoh, K.; Misawa, K.; Kuma, K.; Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **2002**, *30*, 3059–3066. https://doi.org/10.1093/nar/gkf436.
9. Edgar, R. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **2004**, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340.
10. Notredame, C.; Higgins, D.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **2000**, *302*, 205–217. https://doi.org/10.1006/jmbi.2000.4042.
11. Floden, E.W.; Tommaso, P.D.; Chatzou, M.; Magis, C.; Notredame, C.; Chang, J.M. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. *Nucleic Acids Research* **2016**, *44*, W339–W343. https://doi.org/10.1093/nar/gkw300.
12. Cserzö, M.; Bernassau, J.M.; Simon, I.; Maigret, B. New Alignment Strategy for Transmembrane Proteins. *Journal of Molecular Biology* **1994**, *243*, 388 – 396. https://doi.org/https://doi.org/10.1006/jmbi.1994.1666.
13. Zambrano-Vega, C.; Oviedo, B.; Villamar-Torres, R.; Botto-Tobar, M.; Barros-Rodríguez, M. An Overview of Multiple Sequence Alignment Methods Applied to Transmembrane Proteins. In Proceedings of the Technology Trends; Botto-Tobar, M.; Pizarro, G.; Zúñiga-Prieto, M.; D'Armas, M.; Zúñiga Sánchez, M., Eds., Cham, 2019; pp. 410–419.
14. Shafrir, Y.; Guy, H.R. STAM: simple Transmembrane Alignment Method. *Bioinformatics* **2004**, *20*, 758–769. https://doi.org/10.1093/bioinformatics/btg482.
15. Forrest, L.R.; Tang, C.L.; Honig, B. On the accuracy of homology modeling and sequence alignment methods applied to membrane proteins. *Biophysical Journal* **2006**, *91*, 508–517. https://doi.org/10.1529/biophysj.106.082313.
16. Zambrano-Vega, C.; Nebro, A.J.; García-Nieto, J.; Aldana-Montes, J.F. M2Align: parallel multiple sequence alignment with a multi-objective metaheuristic. *Bioinformatics* **2017**, *33*, 3011–3017. https://doi.org/10.1093/bioinformatics/btx338.
17. Cedeño-Muñoz, J.; Zambrano-Vega, C.; Nebro, A.J. TM-MSAligner: A Tool for Multiple Sequence Alignment of Transmembrane Proteins. In Proceedings of the Computational Science – ICCS 2024; Franco, L.; de Mulatier, C.; Paszynski, M.; Krzhizhanovskaya, V.V.; Dongarra, J.J.; Sloot, P.M.A., Eds., Cham, 2024; pp. 113–121.
18. Krogh, A.; Larsson, B.; Von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *Journal of molecular biology* **2001**, *305*, 567–580.
19. Tusnády, G.E.; Simon, I. The HMMTOP transmembrane topology prediction server . *Bioinformatics* **2001**, *17*, 849–850. https://doi.org/10.1093/bioinformatics/17.9.849.
20. Yang, J.; Shen, H.B. MemBrain-contact 2.0: a new two-stage machine learning model for the prediction enhancement of transmembrane protein residue contacts in the full chain. *Bioinformatics* **2017**, *34*, 230–238. https://doi.org/10.1093/bioinformatics/btx593.

 doi:10.20944/preprints202510.0109.v1

21. Cai, Y.D.; Zhou, G.P.; Chou, K.C. Support Vector Machines for Predicting Membrane Protein Types by Using Functional Domain Composition. *Biophysical Journal* **2003**, *84*, 3257–3263. https://doi.org/https://doi.org/10.1016/S0006-3495(03)70050-2.
22. Hallgren, J.; Tsirigos, K.D.; Pedersen, M.D.; Armenteros, J.J.A.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. *bioRxiv* **2022**. https://doi.org/10.1101/2022.04.08.487609.
23. Dayhoff, M.; Schwartz, R.; B.C. Orcutt, B. A model of evolutionary change in proteins. *In Atlas of Protein Sequences and Structure* **1978**, *5*, 345–352.
24. Henikoff, S.; Henikoff, J. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **1992**, *89*, 10915–10919.
25. Frishman, D. *Structural bioinformatics of membrane proteins*; Springer, 2010.
26. Jones, D.; Taylor, W.; Thornton, J. A mutation data matrix for transmembrane proteins. *FEBS letters* **1994**, *339*, 269–275.
27. Ng, P.C.; Henikoff, J.G.; Henikoff, S. PHAT: a transmembrane-specific substitution matrix. *BIOINFORMATICS* **2000**, *16*, 760–766.
28. Müller, T.; Rahmann, S.; Rehmsmeier, M. Non-symmetric score matrices and the detection of homologous transmembrane proteins. *Bioinformatics* **2001**, *17*, S182–S189. https://doi.org/10.1093/bioinformatics/17.suppl_1.S182.
29. Jimenez-Morales, D.; Adamian, L.; Liang, J. Detecting remote homologues using scoring matrices calculated from the estimation of amino acid substitution rates of beta-barrel membrane proteins. In Proceedings of the 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Aug 2008, pp. 1347–1350. https://doi.org/10.1109/IEMBS.2008.4649414.
30. Pirovano, W.; Abeln, S.; Feenstra, K.A.; Heringa, J. Multiple alignment of transmembrane protein sequences **2010**. pp. 103–122.
31. Durillo, J.J.; Nebro, A.J. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software* **2011**, *42*, 760–771.
32. Nebro, A.J.; Durillo, J.J.; Vergne, M. Redesigning the jMetal Multi-Objective Optimization Framework. In Proceedings of the Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation, New York, NY, USA, 2015; GECCO Companion '15, p. 1093–1100. https://doi.org/10.1145/2739482.2768462.
33. Deb, K.; Pratap, A.; Agarwal, S.; Meyarivan, T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* **2002**, *6*, 182 – 197. https://doi.org/10.1109/4235.996017.
34. Madeira, F.; Pearce, M.; Tivey, A.R.N.; Basutkar, P.; Lee, J.; Edbali, O.; Madhusoodanan, N.; Kolesnikov, A.; Lopez, R. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Research* **2022**, *50*, W276–W279. https://doi.org/10.1093/nar/gkac240.
35. Bahr, A.; Thompson, J.D.; Thierry, J.C.; Poch, O. BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucleic acids research* **2001**, *29*, 323–326.
36. Pándy-Szekeres, G.; Munk, C.; Tsonkov, T.M.; Mordalski, S.; Harpsøe, K.; Hauser, A.S.; Papadatos, G.; Siderius, M.; Egebjerg, J.; Jensen, H.H.; et al. GPCRdb in 2023: state-specific structure models using AlphaFold2 and new ligand resources. *Nucleic Acids Research* **2023**, *51*, D368–D376. https://doi.org/10.1093/nar/gkac970.