Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Propedia 26: An expanded and updated database of protein-peptide interactions for machine learning applications

Diego Mariano

diego@dcc.ufmg.br

Universidade Federal de Minas Gerais – UFMG

Adenilson Arcanjo

Instituto Federal de Educação, Ciência e Tecnologia do Ceará

Leonardo Henrique Silva

Universidade Federal de Minas Gerais – UFMG

Milenna Machado Pirovani

Universidade Federal de Minas Gerais – UFMG

Leandro Morais

Universidade Federal de Minas Gerais – UFMG

Luana Luiza Bastos

Universidade Federal de Minas Gerais – UFMG

Rafael Pereira Lemos

Universidade Federal de Minas Gerais – UFMG

Pedro Martins

Universidade Federal de Minas Gerais – UFMG

Raquel Cardoso de Melo-Minardi

Universidade Federal de Minas Gerais – UFMG

---

data-descriptor

Keywords:

Posted Date: February 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8349237/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Propedia is a curated database of protein-peptide interactions. In this new release, we expanded the dataset to more than four times the size of the previous version and introduced a comprehensive set of new annotations. We systematically analyzed protein-peptide complexes to derive predicted physicochemical properties, solvent-accessible surface areas, interaction energies, interface residues, and interchain contacts. In addition, we applied machine learning models to predict the potential therapeutic functions of the peptides. Propedia 26 comprises 98,779 entries, including 78,148 protein–peptide pair complexes (pep-pro) and 20,631 multi-protein complexes featuring a single peptide interacting with one or more protein chains (multipro). We anticipate that this expanded and enriched resource will support the discovery and development of peptide-based drugs and biotechnological products. Propedia 26 is freely accessible at https://bioinfo.dcc.ufmg.br/propedia26.

Background & Summary

Peptides are molecules composed of 2 to 50 amino acid residues connected by peptide bonds [1]. In cells, they play several essential roles in regulation and cellular communication, acting as neurotransmitters, hormones, antimicrobial agents, and immune signaling molecules. Furthermore, they influence various processes such as growth, wound healing, and defense against pathogens. It is estimated that peptides mediate 15% to 40% of molecular interactions between proteins [2]. Recently, peptides have gained great pharmaceutical and biotechnological importance due to their low toxicity, high structural flexibility, and high interaction specificity [3]. Therefore, they have been the target of several studies aimed at developing new drugs and targeted therapies, as well as producing new biotechnological products [4, 5]. However, screening of peptides with biotechnological potential still relies heavily on high-cost, low-scalability experimental approaches. Hence, computational methods have emerged as an innovative alternative for virtual peptide screening [6]. In this context, the creation of peptide databases has become essential for such studies.

Propedia is a database of protein-peptide interactions. The first version of Propedia was released in early 2021 [7]. Propedia consists of data retrieved from the experimental repository Protein Data Bank (PDB), which are then processed and curated [8]. These data were then filtered, and various analyses were performed to extract new features. Initially, Propedia only handled entries consisting of unique protein-peptide pairs. Propedia v1 contained 19,813 protein-peptide complexes clustered by three methods: sequence-based, interface-based, and binding-site-based.

Sequence-based clusters were constructed using the Hammock tool v1.2 [9]. The Hammock pipeline identified 3,495 unique sequences, categorizing them into 771 sequence clusters and 1,074 unique sequences (singletons), resulting in a total of 1,845 peptide sequence clusters. Interface clusters were constructed using the MUSTANG tool [10], which performs multiple structure alignments. MUSTANG detected 535 clusters, in addition to 1,356 singletons, resulting in 1,891 protein-peptide interfaces. Lastly, clusters were also constructed based on binding-site similarities using the ProBiS algorithm [11].

ProBIS employs a local alignment algorithm to identify similar binding sites in proteins with distinct folds. To achieve this, it utilizes geometric and functional groups to detect three-dimensional patterns of physicochemical properties on their surfaces. Ultimately, 521 binding site clusters were detected, in addition to 945 singletons, totaling 1,466 different binding sites. Among the main features presented by Propedia v1, the binding-site search engine stands out. In practice, if a user wants to find peptides that bind to a specific known binding site, they would enter the PDB code, the chain, and the residues that comprise the site into the interface. Propedia would then use the ProBIS algorithm and its database of binding-site-based clusters to detect experimentally resolved peptides that potentially bind to that binding site.

The second version of Propedia was released in late 2023 [12]. It presented 49,300 protein-peptide complexes. Among the new features, it introduced a set of subdatasets classified according to the protein class extracted from the Protein Data Bank. Among the classifications, we can mention: antimicrobial, viral, enzyme, membrane, hormone, and plant. Furthermore, structural signatures were calculated for each protein-peptide complex, receptor, and peptide individually using the aCSM algorithm [13]. Structural signatures are mathematical representations used to encode three-dimensional structural information in a one-dimensional vector space [14].

In this paper, we present Propedia 26, the third version of the protein-peptide interaction database. The version nomenclature has been changed to reflect the year of release. Propedia 26 contains 4 times as many entries as the first version and introduces several new features to characterize protein-peptide complexes better. Furthermore, Propedia 26 introduces calculations for each entry of seven types of interchain contacts: hydrogen bonds, attractive bonds, repulsive bonds, hydrophobic bonds, aromatic bonds, salt bridges, and disulfide bonds. We also offer several new features that may be beneficial for classification tasks utilizing machine learning algorithms.

Data Records

Briefly, Propedia 26 contains 98,779 entries, with 78,148 protein-peptide pair complexes (pep-pro) and 20,631 complexes composed of a single peptide interacting with two or more proteins (multipro). Figure 1 summarizes the steps taken to collect the data that make up Propedia 26. First, the data were collected from the PDB database. We selected structures with peptides close to proteins (distance ≤ 6Å). We separated the protein-peptide structure pairs and calculated a set of features using various approaches.

For each entry, Propedia 26 provides a comprehensive set of structural, physicochemical, and functional annotations. Complex-level characteristics include chain identifiers, structural descriptions, sequence length (in residues), molecular weight (Da), isoelectric point (pI), instability and aliphatic index, GRAVY (Grand Average of Hydropathy), percentage of hydrophobic residues, number of positively and negatively charged residues, atomic formula, molar extinction coefficient ($M^{-1}cm^{-1}$), and full sequence information. These characteristics were calculated using the ProtParam library [15]. The dataset also

lists structurally and sequentially similar complexes, facilitating comparison and similarity-based analyses.

Previous versions of Propedia only considered protein-peptide interaction pairs. For example, consider the PDB entry 1A1R, which contains three chains (A, B, and C). Chain C belongs to a peptide that interacts simultaneously with the protein in chains A and B. By default, the Propedia pipeline saves two entries: 1A1R-C-A (peptide C interacting with the protein chain A) and 1A1R-C-B (peptide C interacting with the protein chain B). In this new version of Propedia, we decided to include a new type of entry that groups all chains interacting with the same peptide. We refer to this dataset as Propedia Multipro. In this case, the entry 1A1R-C would contain the C chains of the peptide and the A and B chains of proteins that are interacting with it. We did this for all entries in Propedia 26. In total, we collected 20,631 Multipro complexes.

Propedia 26 further reports the predicted functional profiles of each peptide, indicating the probability of exhibiting anti-angiogenic (AAP), antibacterial (ABP), anticancer (ACP), anti-inflammatory (AIP), quorum-sensing (QSP), or surface-binding (SBP) activity. The original Propedia v1 annotation schema, including Binding Site, Interface, and Sequence classes, is fully retained.

A detailed contact surface analysis is also provided. The Accessible Surface Area (ASA), expressed in Å², quantifies the portion of the molecule exposed to solvent (typically water). The Buried Surface Area (BSA) represents the region effectively engaged at the binding interface. Additionally, we report the Buried Peptide surface Percentage (BPP%), indicating the fraction of the peptide that is not exposed to the solvent. All ASA- and BSA-related measurements were obtained using NACCESS [16]. Additionally, Propedia 26 also includes predictions of protein-peptide interaction energy, estimating binding affinity (kcal/mol) and dissociation constant (M) at 25°C using the PRODIGY algorithm [17].

Finally, the database identifies all interface residues, defined as protein residues within 6 Å of the peptide, and provides an atom-level catalog of interchain contacts. The computed contact types include hydrogen bonds, attractive and repulsive interactions, hydrophobic contacts, π-π stacking, salt bridges, and disulfide bonds. All contacts were calculated using the COCαDA algorithm [18].

A description of the table of entries in Propedia is provided in Supplementary Table S1.

Technical validation

Data quality was assessed through random sampling of the dataset. A subset of entries was randomly selected, and the values presented in the Propedia interface were manually cross-checked against their respective primary data sources. To evaluate the descriptors, we consulted the Protein Data Bank page corresponding to the entry. For the analysis of physicochemical parameters, we collected the peptide sequence and used the ProtParam web tool (https://web.expasy.org/protparam). To evaluate the interaction energy, we consulted the Prodigy web tool, available at https://rascar.science.uu.nl/prodigy. The list of intra-chain contacts was verified using the COCαDa web tool, available at

https://bioinfo.dcc.ufmg.br/cocada-web. The NACCESS surface area data were evaluated using in-house Python scripts. In addition, the Orange Data Mining tool was used to assess the feature statistics and detect distribution, mean, mode, median, dispersion, minimum and maximum values, and missing values.

Missing data in Propedia 26 arise from isolated issues during the execution of specific computational tools, as detailed in the Methods section. For example, the presence of non-canonical amino acids or high cutoff distances between chains (by default, Propedia defines the maximum distance between chains as 6Å, but other tools may require smaller distances) can lead to errors in the tools' calculations. Overall, data completeness is high across the dataset. Protein surface area measurements are missing for 256 entries (< 1% of the dataset), while predicted therapeutic properties are unavailable for 1,875 structures (2%). Interaction energy predictions could not be computed for 5,010 entries (6%), primarily due to limitations in structural quality or contact detection required by the prediction algorithm. Predicted physicochemical properties for protein chains are missing for 41 entries (< 1%). Importantly, all entries include experimentally derived complex-level information obtained from the Protein Data Bank, and 100% of peptide chains contain predicted physicochemical descriptors. These statistics indicate that missing values are limited, well-characterized, and confined to specific derived features, ensuring the overall reliability and reusability of the dataset. The data report is available on the supplementary material's GitHub repository.

Usage Notes

Propedia 26 web interface

Propedia 26 has a user-friendly web interface designed to simplify navigation and search of structures. There are three entry search mechanisms: (i) sequence-based search (uses BLAST to perform local alignments with parameters optimized for peptide search); (ii) binding site-based search (uses ProBis to detect structures with cavities similar to the region indicated in the target protein by users); and (iii) search based on Propedia ID, PDB ID, or entry metadata (Fig. 2 - top). Each entry in Propedia has its own page with information such as a description of the entry, a presentation of physicochemical parameters detected *in silico*, and an interactive visualization of the 3D structure (Fig. 2 - below).

The entry page also presents similar structures, predicted characteristics for therapeutic peptides, and the clusters inherited from Propedia v1 (based on sequence, interface, and binding site; Fig. 3A). Furthermore, information on protein-peptide interactions is calculated using different software. Solvent-accessible area is predicted using NACCESS [16]. We display information about the solvent-exposed area for the complex, only the peptide, and only the protein (Fig. 3B). In addition, we show the buried areas of the peptide (BPepA) and the protein (BProA), corresponding to the regions that mediate the protein-peptide interaction. Note that, mostly, these regions occupy different area values in the protein and in the peptide. Therefore, we use the average of these values to indicate the area of the binding interface (*i.e.*, buried surface area - BSA). We also report the percentage of the peptide that is buried

(BPP%). The higher the BPP% value, the more deeply embedded the peptide is, and therefore, the stronger the bond tends to be. However, it should be noted that the peptide may form bonds with other protein chains (see the complex's multipro page for more details).

Interaction energy, binding affinity, and dissociation constant are predicted using Prodigy (Fig. 3B). We also list the residues in the protein interface (Fig. 3C). Finally, interchain contacts were calculated using the COCαDA-CLI (Fig. 3D). The contact list table is integrated into the 3D structure visualization. Clicking on a listed contact displays the interacting residues as sticks (Fig. 3E).

Case study: Searching for peptides that inhibit the SARS-CoV-2 Mpro protein

In this section, we illustrate the application of the binding site search algorithm to identify peptides that interact with the SARS-CoV-2 Mpro protein. This section is not intended to provide an in-depth biological analysis of the case studies or predictions; instead, it aims to offer illustrative examples of the database entries and their potential use cases.

SARS-CoV-2 remains a global concern due to its complex pathogenesis and its potential to cause severe disease, particularly when viral replication is not effectively controlled in target tissues [19–21]. Mpro (Main Protease), also called 3CLpro, is one of the most essential enzymes for the SARS-CoV-2 life cycle. It plays a central role in viral replication, which is why it has become one of the most critical targets for antiviral drug development.

The Propedia binding site search requires as input a PDB code of the target protein, the chain, and a list of residue codes. The tool then uses the ProBis algorithm to search for structures with binding sites that have conformations similar to those of the input structure.

The binding site used in our search was defined according to recent structural studies [22], focusing on the key residues that form the catalytic region of Mpro. Propedia 26 identified 180 protein-peptide complexes matching the defined binding site, whereas Propedia v1 returned only 34. This increase reflects the expansion of the database, in particular, due to the large number of recent studies on SARS-CoV-2. This highlights the importance of updating the Propedia database. Figure 4 presents the top three peptides obtained in our search, ranked by Alignment Score and RMSD.

Methods

Data collection

We collected all structures with a maximum of 10,000 amino acid residues from the Protein Data Bank (PDB) [8] in CIF format on September 8, 2025 (n = 236,616). Then, we filtered all complexes performing protein-peptide interactions using Python scripts. We defined a complex as exhibiting a protein–peptide

interaction when at least one atom from the protein and one atom from the peptide were separated by a distance of 6 Å or less. Additionally, we extracted all combinations of chain pairs where one chain had between 2 and 50 amino acids, while the other had more than 50 amino acids. In total, we extracted 78,148 protein-peptide complexes.

Furthermore, we created a sub-dataset comprising all protein chains within a maximum distance of 6 Å from the peptide. We called this the Multipro dataset. In this case, each entry point of the multipro is composed of one peptide interacting with two or more protein chains. In total, we obtained 20,631 multipro entries.

Clustering and Redundancy Detection

To detect redundancy in protein-peptide complexes, we concatenated the protein and peptide sequences and grouped identical sequence combinations. In total, we detected 51,082 unique protein-peptide complexes.

Then, we collected all peptide sequences containing only canonical amino acids and grouped those that were 100% identical. In total, we extracted 17,509 unique sequences composed only of the 20 canonical amino acids. We refer to this peptide dataset as the canonical non-redundant (CNR) dataset.

Clusters described in previous versions of Propedia have been recalculated using Python scripts. Other classification information was obtained from the PDB file using scripts developed with the Biopython library (Bio.PDB package) [23, 24].

Feature extraction

For each protein-peptide complex, physicochemical properties were obtained using ProtParam [15] embedded in the Biopython library v1.85 [23]. The collected information included the chain identifier, description, sequence length (number of residues), molecular weight (Da), isoelectric point (pI), instability index, aliphatic index, GRAVY (grand average of hydropathicity), hydrophobicity (%), total number of positively and negatively charged residues, atomic formula, number of atoms, extinction coefficient (with and without disulfide bonds), and the full amino acid sequence.

Protein-peptide interactions

Protein interaction interface residues were calculated using Python scripts. The IDs of proteins' residues at a distance ≤ 6 Å from the peptide were annotated.

The solvent-accessible surface area (ASA) was calculated using NACCESS v2.1.1 [16]. For each protein–peptide complex, the total ASA of the complex, the changes in ASA for the protein and the peptide upon binding (ΔASA), and the buried surface area (BSA) were computed and reported in square angstroms

(Å²). The Buried Surface Area (BSA), which represents the region effectively engaged at the binding interface, is computed as:

$$BSA = \frac{ASA_{protein} + ASA_{peptide} - ASA_{complex}}{2}$$

Furthermore, we calculated the buried area of peptides (BPepA) and proteins (BProA). Finally, we calculated the buried peptide percentage (BPP%) using the formula:

$$BPP\,(\%) = \frac{100 \times BPepA}{ASA_{peptide}}$$

The interaction energy of each protein–peptide complex was calculated using PRODIGY v.2.4.0 [17]. For each complex, the total number of interfacial contacts (categorized by amino acid type), the predicted binding affinity (kcal/mol), and the predicted dissociation constant (M) at 25°C were computed.

Chain contacts were calculated using the COCαDA command-line tool v1.0 [18, 25]. For each complex, COCαDA calculated hydrogen bonds, attractive and repulsive interactions, hydrophobic interactions, aromatic interactions, disulfide bonds, and salt bridges.

Structural signature

For each three-dimensional structure of unique peptides composed only of canonical amino acids (CNR dataset), we calculated structural signatures using the SIGNA library (https://github.com/LBS-UFMG/signa). Structural signatures convert the 3D structure into one-dimensional vectors that can be used in machine learning tasks. For the calculation of structural signatures, we used the aCSM-ALL algorithm [13] with a maximum distance parameter of 10 Å and a step of 0.2 Å. These parameters were based on previous studies [12, 14, 26–28].

Sequence signature

Furthermore, we extracted features based on the sequences using the iFeature tool (https://github.com/Superzchen/iFeature) [29]. We named the extracted sequence features as sequence signatures. Sequence-based signatures can be crucial for machine learning tasks that involve predictions where three-dimensional structures are unavailable. We selected the 10 following descriptors: AAC, DPC, DDE, GAAC, GDPC, GTPC, CTDC, CTDT, CTDD, and CTriad.

The Amino Acid Composition (AAC) descriptor represents the relative frequency of each of the 20 amino acids in a peptide sequence. The Dipeptide Composition (DPC) descriptor extends this idea by quantifying the frequencies of all 400 possible dipeptide pairs, thereby capturing short-range order information. The Dipeptide Deviation from Expected mean (DDE) descriptor measures how much the observed dipeptide frequencies deviate from values expected based on individual amino acid compositions, reflecting pairwise dependency patterns. The Grouped Amino Acid Composition (GAAC) descriptor simplifies the sequence by clustering amino acids into physicochemical groups and

computing their overall proportions. Similarly, the Grouped Dipeptide Composition (GDPC) and Grouped Tripeptide Composition (GTPC) descriptors compute the frequencies of dipeptides or tripeptides based on predefined amino-acid groups rather than individual residues, reducing dimensionality while retaining key biochemical patterns. The Composition Transition Distribution Composition (CTDC) descriptor summarizes the fraction of residues belonging to different physicochemical categories, whereas CTD Transition (CTDT) quantifies how often the sequence switches between these categories. CTD Distribution (CTDD) further describes the positional distribution of residues from each category along the sequence. Finally, the Conjoint Triad (CTriad) descriptor groups amino acids into seven classes and measures the frequency of all possible triads of these groups, capturing local physicochemical environments.

Finally, iFeature returns a vector of 1248 features for each sequence used as input. In this work, we refer to this vector as a sequence signature.

Therapeutic peptides classification

We classified each peptide in Propedia into six classes according to their predicted therapeutic properties: Anti-Angiogenic (AAP), Antibacterial (ABP), Anticancer (ACP), Anti-Inflammatory (AIP), Quorum Sensing (QSP), and Surface Binding (SBP). The classification was performed using machine learning models with Orange Data Mining [30]. To build the models, we used sequence-based signatures and attempted to reproduce the methodology described in the article by Rodrigues et al. [31]. Details of the methodology are available in our GitHub's supplementary material repository (see Data Availability section).

User-friendly interface development

The web interface was built using the same approaches used in [32–35]. The CodeIgniter 4 framework was used for the back-end. The Bootstrap and DataTables libraries were used in the front-end application. Interactive protein visualizations were built using the 3Dmol.js library [36]. Contact map graphs were built using the Chat.js library. Figures used to illustrate the sections were constructed using the ChimeraX tool [37].

The sequence-based search system was built using BLAST [38, 39]. To enable searches for peptide sequences, the following parameters were used: "-word_size 2 -task blastp-short -seg no -evalue 100000". For protein sequence searches, only the parameter "-word_size 3" was employed.

To search for similar binding sites, the Propedia back-end uses the ProBiS tool [11]. The ProBiS algorithm detects local structural similarities by comparing the three-dimensional surface of the protein's binding site with the 3D surfaces of proteins stored in the Propedia-26 database. To reduce search costs, we selected only non-redundant structures. Additionally, some structures exhibited errors during the extraction of the peptide interaction surface. In total, 32,695 protein-peptide complex surfaces are used in the search for similar binding sites.

Declarations

Competing interests

The authors declare no competing interests.

Funding

This work was supported by grants from the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, projects APQ-02690-22 and APQ-01834-21) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, project 440307/2022-8). This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Author Contribution

P.M., D.M., and R.C.M.M contributed to the conceptualization; D.M. collected the data and developed the web tool; L.H.S., M.M.P, and L.M. performed the case studies; M.M.P wrote the documentation; L.B. performed the surface analysis; A.A. performed the analysis of interaction energy; R.L. performed the contact analysis; D.M. performed the analysis of physical-chemical parameters; R.C.M.M. acquired funds and supervised the research. D.M. prepared the manuscript with the assistance and feedback from all other co-authors.

Acknowledgement

The authors would like to express their gratitude to the research funding agencies: CAPES, CNPq, and FAPEMIG. The authors thank Felipe Conceição for his valuable contributions.

Data availability

The web tool can be found at https://bioinfo.dcc.ufmg.br/propedia26. The dataset and data descriptors can be downloaded at https://bioinfo.dcc.ufmg.br/propedia26/public/download. Full supplementary material is available at https://github.com/LBS-UFMG/propedia26-sm. Data descriptors are available at Supplementary Table S1.

Code availability

The source code of the web tool can be found at: https://github.com/LBS-UFMG/propedia26. The Orange Data Mining machine learning models are available in our GitHub supplementary material repository: https://github.com/LBS-UFMG/propedia26-sm.

References

1. Sauvestre, C., Zagury, J.-F. & Langenfeld, F. Comparative Analysis of Deep Learning-Based Algorithms for Peptide Structure Prediction. *Proteins Struct. Funct. Bioinforma.* n/a, (2025).
2. Neduva, V. *et al.* Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.* 3, e405 (2005).
3. Lee, A. C.-L., Harris, J. L., Khanna, K. K. & Hong, J.-H. A comprehensive review on current advances in peptide drug development and design. *Int. J. Mol. Sci.* 20, 2383 (2019).
4. Wang, L. *et al.* Therapeutic peptides: current applications and future directions. *Signal Transduct. Target. Ther.* 7, 48 (2022).
5. Cabral, H. Peptides: Molecular and Biotechnological Aspects. *Biomolecules* 11, 52 (2021).
6. Tripathi, N. M. & Bandyopadhyay, A. High throughput virtual screening (HTVS) of peptide library: Technological advancement in ligand discovery. *Eur. J. Med. Chem.* 243, 114766 (2022).
7. Martins, P. M. *et al.* Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. *BMC Bioinformatics* 22, 1 (2021).
8. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242 (2000).
9. Krejci, A., Hupp, T. R., Lexa, M., Vojtesek, B. & Muller, P. Hammock: a hidden Markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. *Bioinforma. Oxf. Engl.* 32, 9–16 (2016).
10. Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J. & Lesk, A. M. MUSTANG: a multiple structural alignment algorithm. *Proteins* 64, 559–574 (2006).
11. Konc, J. & Janežič, D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* 26, 1160–1168 (2010).
12. Martins, P. *et al.* Propedia v2.3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. *Front. Bioinforma.* 3, (2023).
13. Pires, D. E. V., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F. & Meira, W. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinforma. Oxf. Engl.* 29, 855–861 (2013).
14. Mariano, D. *et al.* A Computational Method to Propose Mutations in Enzymes Based on Structural Signature Variation (SSV). *Int. J. Mol. Sci.* 20, (2019).
15. Gasteiger, E. *et al.* Protein identification and analysis tools on the ExPASy server. in *The proteomics protocols handbook* 571–607 (Springer, 2005).
16. Hubbard, S. J. & Thornton, J. naccess. *Dept Biohem Mol Biol UCL* (1993).
17. Vangone, A. & Bonvin, A. M. Contacts-based prediction of binding affinity in protein–protein complexes. *elife* 4, e07454 (2015).
18. Lemos, R., Mariano, D., Silveira, S. & Melo-Minardi, R. *COCαDA - Large-Scale Protein Interatomic Contact Cutoff Optimization by Cα Distance Matrices*. 70 (2024). doi:10.5753/bsb.2024.245545.

19. Ibrahim, M. *et al.* Why is the Omicron main protease of SARS-CoV-2 less stable than its wild-type counterpart? A crystallographic, biophysical, and theoretical study of the free enzyme and its complex with inhibitor 13b-K. 2024.03.04.583178 Preprint at https://doi.org/10.1101/2024.03.04.583178 (2024).
20. Lamers, M. M. & Haagmans, B. L. SARS-CoV-2 pathogenesis. *Nat. Rev. Microbiol.* 20, 270–284 (2022).
21. Hashemi, Z. S. *et al.* In silico Approaches for the Design and Optimization of Interfering Peptides Against Protein–Protein Interactions. *Front. Mol. Biosci.* 8, (2021).
22. Chaves Carvalho, F. *et al.* A hybrid evolutionary and structural method for AI-guided peptide inhibitor design using AlphaFold and Rosetta. *Sci. Rep.* https://doi.org/10.1038/s41598-025-28061-y (2025) doi:10.1038/s41598-025-28061-y.
23. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma. Oxf. Engl.* 25, 1422–1423 (2009).
24. Hamelryck, T. & Manderick, B. PDB file parser and structure class implemented in Python. *Bioinformatics* 19, 2308–2310 (2003).
25. Lemos, R. P., Mariano, D., Silveira, S. de A. & de Melo-Minardi, R. C. COCαDA-A Fast and Scalable Algorithm for Interatomic Contact Detection in Proteins Using Cα Distance Matrices. *Front. Bioinforma.* 5, 1630078 (2025).
26. Moreira, E. U. *et al.* Predicting Mutation-Driven Changes in the SARS-CoV-2 Spike Protein Using Structural Signatures and Neural Networks. in 167–178 (SBC, 2024).
27. Mariano, D. *et al.* Using Machine Learning and Graph-Based Signatures to Evaluate Normal and Decreased Function CYP2D6 Haplotypes. in 90–105 (Springer, 2025).
28. Cioletti, A. G., Lemos, R. P., dos Santos, L. M., Mariano, D. & de Melo-Minardi, R. C. Using graph-based structural signatures and machine learning algorithms for molecular docking assessment of histone deacetylases and small ligands. in 223–234 (SBC, 2024).
29. Chen, Z. *et al.* iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502 (2018).
30. Demšar, J., Zupan, B., Leban, G. & Curk, T. Orange: From experimental machine learning to interactive data mining. in 537–539 (Springer, 2004).
31. Rodrigues, C. H. M., Garg, A., Keizer, D., Pires, D. E. V. & Ascher, D. B. CSM-peptides: A computational approach to rapid identification of therapeutic peptides. *Protein Sci.* 31, e4442 (2022).
32. Silva, M. F. M. *et al.* Proteingo: Motivation, user experience, and learning of molecular interactions in biological complexes. *Entertain. Comput.* 29, 31–42 (2019).
33. Fassio, A. V., Santos, L. H., Silveira, S. A., Ferreira, R. S. & de Melo-Minardi, R. C. nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2019.2892099 (2019) doi:10.1109/TCBB.2019.2892099.

34. Barroso, J. R. M. S. *et al.* Proteus: An algorithm for proposing stabilizing mutation pairs based on interactions observed in known protein 3D structures. *BMC Bioinformatics* 21, 275 (2020).
35. Pimentel, V. *et al.* VTR: a web tool for identifying analogous contacts on protein structures and their complexes. *Front. Bioinforma.* 1, 28 (2021).
36. Rego, N. & Koes, D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics* 31, 1322–1324 (2015).
37. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Sci.* 32, e4792 (2023).
38. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410 (1990).
39. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421 (2009).

Figures

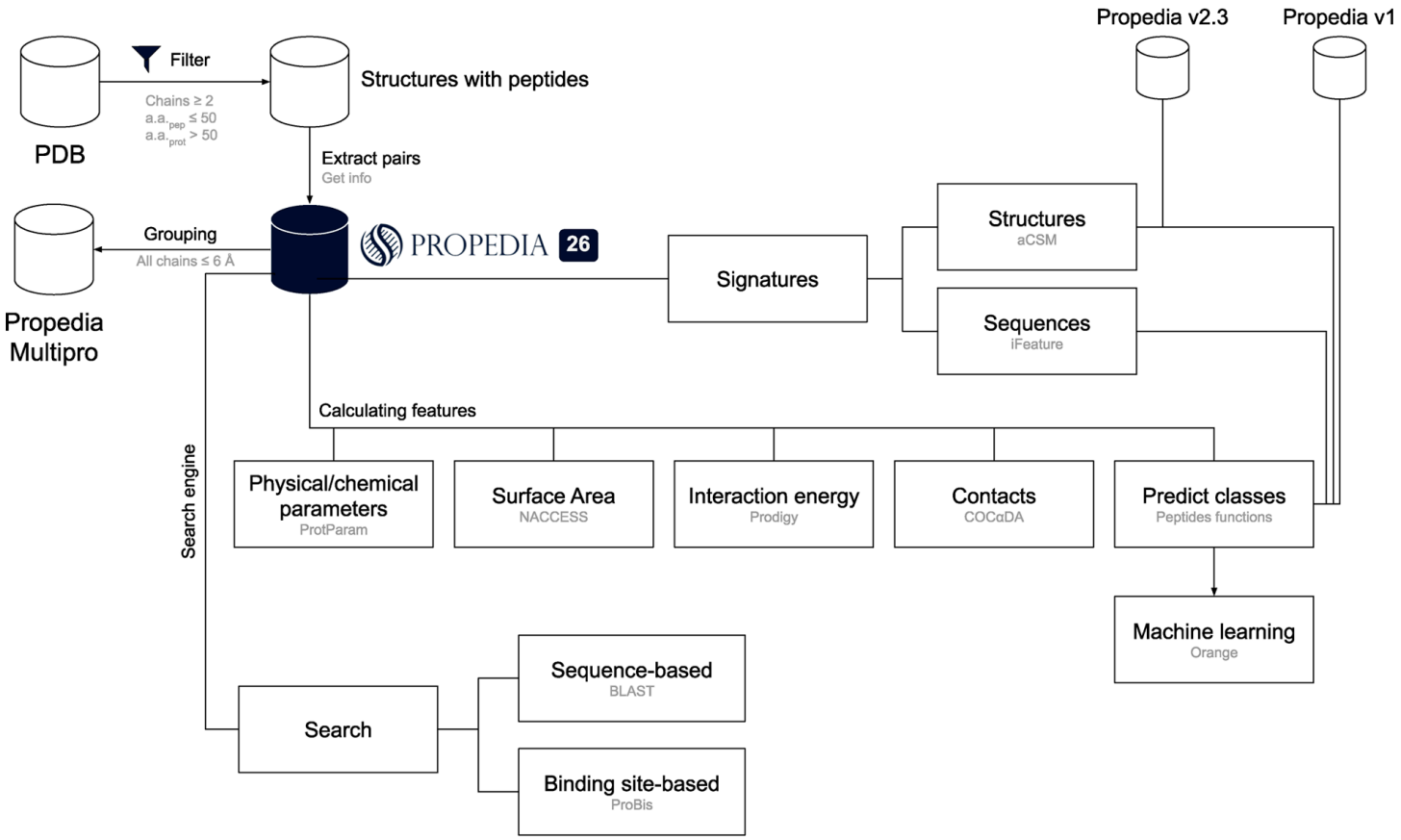


Figure 1

Propedia 26 overview. 3D structures were collected from the PDB database. Complexes containing protein-peptide structures were filtered. Then, we predicted and calculated several features, such as signatures (based on sequence and structures), physical/chemical parameters, surface area, interaction

energy, interatomic contacts, and grouped structures based on their function. The search engine implemented in the interface was developed using a sequence approach (using BLAST) and a binding site approach (using ProBis).

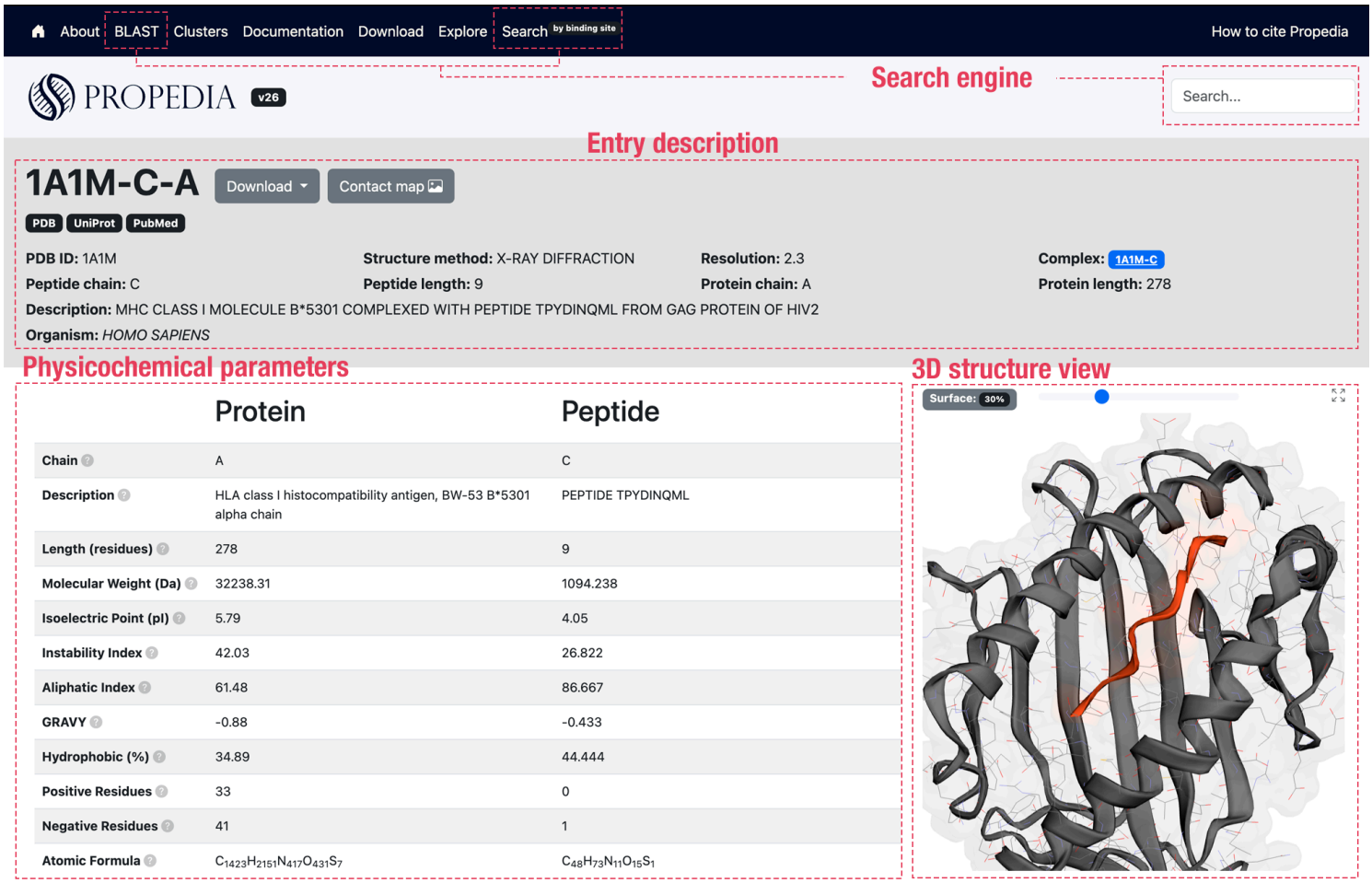


Figure 2

Page from a Propedia entry (ID: 1A1M-C-A). At the top, we can see the three search engines, accessible on any Propedia page. Below, we can see entry details, such as entry description, physicochemical parameters, and interactive 3D structure visualization.

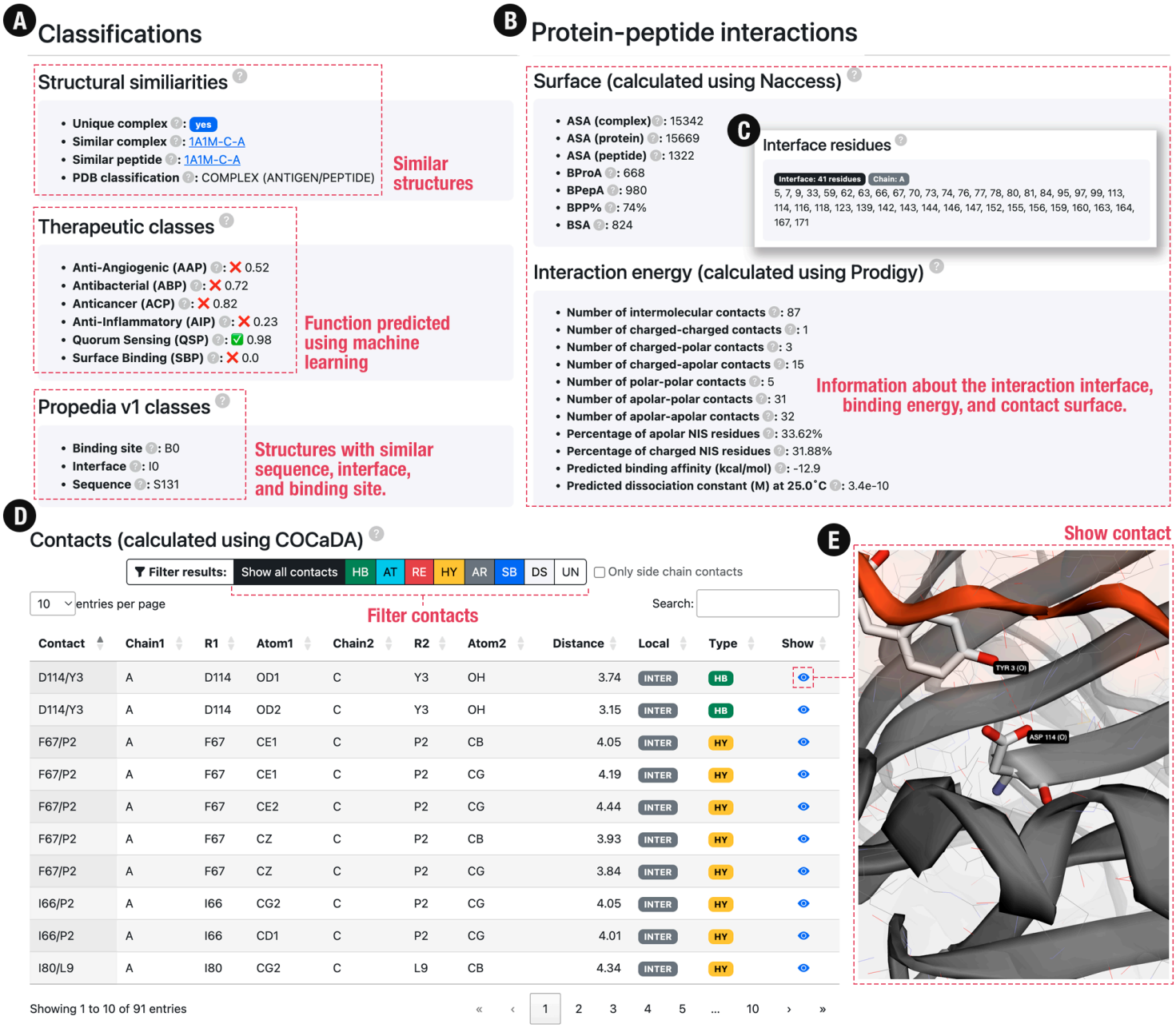


Figure 3

Predicted functions and protein-peptide interaction data are shown in the Entry page (ID: 1A1M-C-A). (A) Similar structures and classifications; (B) Surface and protein-peptide interactions; (C) Interface protein residues; (D) Protein-peptide contacts; and (E) Interactive 3D structure view of the complex. In this example, we are highlighting the contact between D114 and Y3 (hydrogen bond).

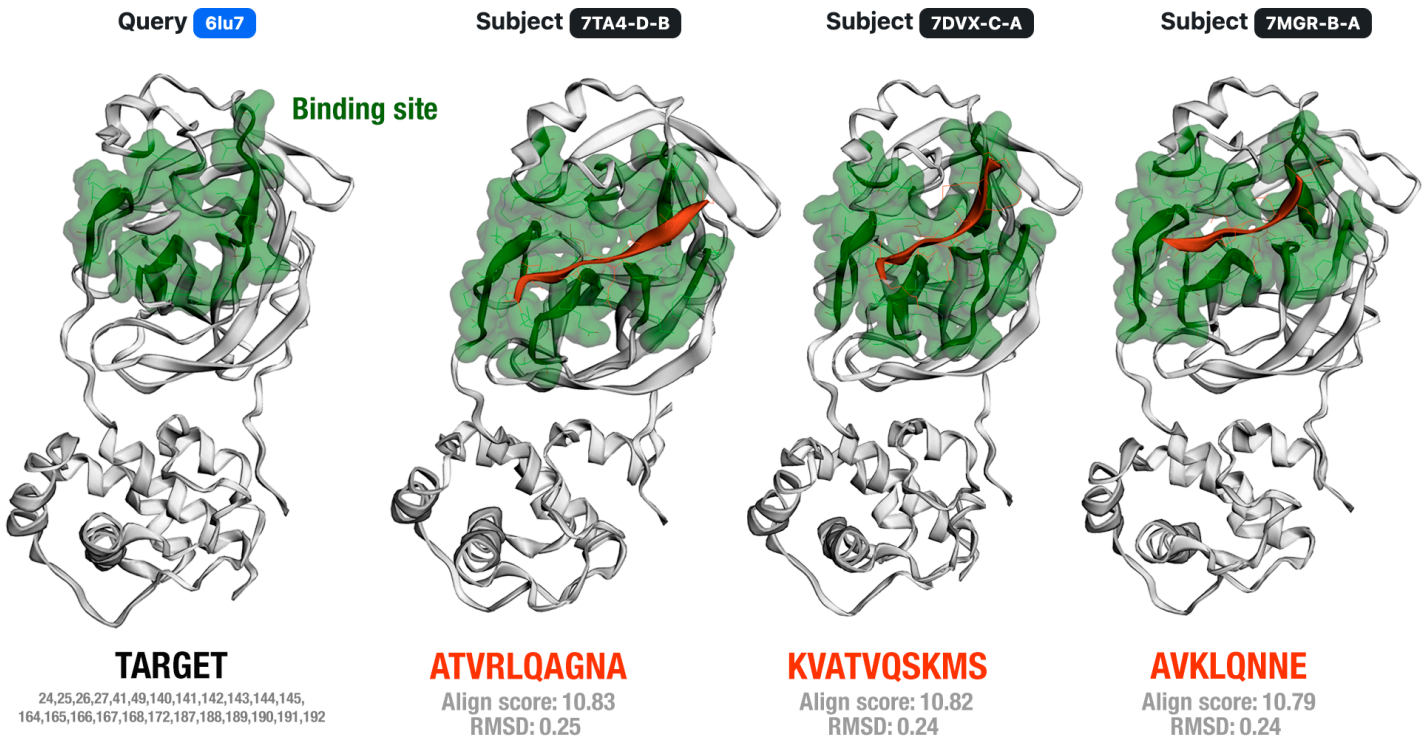


Figure 4

Identification of inhibitory peptides for SARS-CoV-2 Mpro. The target structure (left) highlights the binding site surface in green (PDB ID: 6LU7). Residues that compound the binding site are: 24, 25, 26, 27, 41, 49, 140, 141, 142, 143, 144, 145, 164, 165, 166, 167, 168, 172, 187, 188, 189, 190, 191, 192. The search returned complexes with high structural compatibility, represented by the three peptides with more compatibility to the binding site: ATVRLQAGNA, KVATVQSKMS, and AVKLQNNE. The orange ribbons illustrate how these ligands fit into the active site, illustrating how the binding-site similarity search retrieves peptides with structural compatibility to a target site based on low RMSD values and high Alignment Score values.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterialv2.pdf