Posters

P1

ISOLATION AND FUNCTIONAL ANALYSIS OF PHAGE-DISPLAYED ANTIBODY FRAGMENTS TARGETING THE STAPHYLOCOCCAL SUPERANTIGEN-LIKE PROTEINS

Ida Alanko[1], Rebecca Sandberg[1], Eeva-Christine Brockmann[2], Carla J. C. de Haas[3], Jos A. G. van Strijp[3], Urpo Lamminmäki[2], Outi M. H. Salo-Ahen[1]

[1]Faculty of Sciences and Engineering, Pharmaceutical Sciences Laboratory (Pharmacy) & Structural Bioinformatics Laboratory (Biochemistry) Turku, Åbo Akademi University, Turku, Finland
[2]Department of Life Technologies, University of Turku, Turku, Finland
[3]Department of Medical Microbiology, University Medical Center Utrecht, Utrecht University, Utrecht, The Netherlands
ialanko@abo.fi

Staphylococcus aureus produces an arsenal of virulence factors that manipulate the immune system helping the bacteria avoid phagocytosis. In this study we are investigating two of these evasion molecules called the Staphylococcal superantigen-like protein 1 and 5 (SSL1 and SSL5). Both SSLs inhibit vital host immune processes and contribute to *S. aureus* immune evasion, e.g. by inhibiting the Matrix metalloproteinase 9 (MMP9), thus limiting chemokine potentiation and neutrophil migration. The aim of this study was to isolate single-chain variable fragment (scFvs) antibodies from synthetic antibody phage libraries, that can recognize SSL1 and SSL5, and that could block the interaction between the SSLs and their respective human targets.

The scFv-antibodies were selected after three rounds of panning against SSL1 and SSL5 and their binding activity to the SSL1 and SSL5 was studied using time-resolved fluorescence-based immunoassay. We obtained altogether 27 unique clones displaying binding activity to the SSL1 and SSL5. The capability of the scFvs to inhibit the SSLs' function was tested various immunoassays including an MMP9 enzymatic activity assay. We were able to show that ten scFvs were able to inhibit the SSL1 or SSL5 in a concentration dependent manner. Some antibodies were able to restore the MMP9 activity fully after incubation with scFv-bound SSLs.

Finally, the structure of the best inhibiting scFv was modeled and used to create putative scFv-SSL-complex models by protein–protein docking. The complex models were subjected to a 100-ns molecular dynamics simulation to assess the possible binding mode of the antibody.

We have demonstrated that by utilizing phage display we are able to isolate antibodies that recognize and inhibit virulence factors. The antibodies found here could be a ground for developing antivirulence factors against *S. aureus* infections to help restore the immune system's capacity and further enable a more efficient clearance of the bacteria.

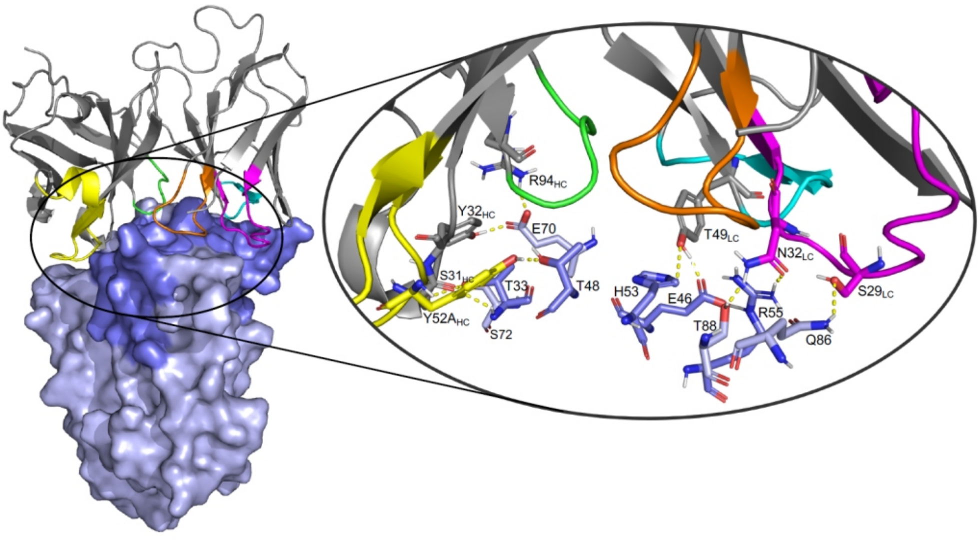


Fig 1. A predicted binding complex of SSL1 with an inhibiting scFv.

P2

DETAILED ANALYSIS OF A THERMOSTABLE PROTEIN-DNA COMPLEX: THE CASE OF Sac7d AS A PROTOTYPE FOR PROTEIN-DNA INTERACTION

Elena Álvarez[1,2], Stéphane Téletchéa[1], Simon Huet[2], Bernard Offmann[1]

[1]Nantes Université, US2B, CNRS, UMR6286 & Affilogic F-44000 Nantes, France
[2]Affilogic SAS, F-44000 Nantes, France
elena.alvarezsanchez@affilogic.com

Sac7d is a 7kDa protein belonging to the class of the small chromosomal proteins from archeon *Sulfolobus acidocaldarius*. Sac7d was discovered in 1974 in Yellowtone National Parks geysers, and studied extensively since then for its remarkable stability at large pH and temperature ranges. Sac7d binds to DNA minor groove by raising its melting temperature, thus protecting DNA from these extreme conditions.

In this study, we analyzed Sac7d-DNA complex using 1µs molecular dynamics simulations to determine which residues contributed most to DNA binding. The interaction energy of the interface was decomposed using Molecular Mechanics Generalized Born Surface Area (MM/GBSA). We determined that more than 10 residues were critical for DNA recognition. The individual contribution of each residue to the binding interface was in agreement with previous documented results. We provide a novel in-depth focus on the DNA energetics as a consequence of its tethering to Sac7d.

P3

METHOD FOR ANALYSIS OF PROTEIN-DNA INTERACTIONS USING PROBABILITY DENSITY MAPS

Daniel Berdár, Bohdan Schneider, Lada Biedermannová

Institute of Biotechnology, Czech Academy of Science, Czech Republic
daniel.berdar@ibt.cas.cz

A plethora of experimental and computational tools are used to improve the understanding of biomolecular structure-function relationships. At the intersection of theoretical and experimental approaches stand data analysis studies that use a large pool of structures available within structural databases, such as the PDB, to show patterns in large ensembles of macromolecular structures.

Here we plan to demonstrate how, using NtC's for local conformational description of nucleic acids, we can construct probability density maps for DNA fragments. Maps are constructed for relevant combinations of DNA building blocks and selected interacting atom/group. These can be further examined or superposed to elucidate guiding patterns of interaction at biomolecular interfaces.

P4

PREDICTION OF DNA HYDRATION BASED ON DATA MINING OF CRYSTALLOGRAPHIC STRUCTURES

Lada Biedermannová, Bohdan Schneider

Institute of Biotechnology, Czech Republic
Lada.Biedermannova@ibt.cas.cz

Water is a critical factor in stabilizing DNA structure and mediating its interactions. In our study, we harness crystallographic data to establish average hydration patterns around biomolecules, including proteins [1,2,3] and nucleic acids [4,5,6]. Our recent focus has been on exploring DNA hydration as a function of its conformation and sequence.

To gain a more comprehensive understanding, we employed a multi-step approach to determine water probability densities around dinucleotide fragments. Beginning with DNA crystal structures containing water molecules, we conducted an extensive analysis of DNA dinucleotides within an ensemble of 2,727 non-redundant DNA chains, encompassing 41,853 dinucleotides and the associated 316,265 first-shell water molecules [6]. We classified dinucleotides based on their 16 sequences and the previously defined structural classes, known as nucleotide conformers (NtCs). From the DNA structures in the data set, we extracted all dinucleotides with associated water molecules. Subsequently, all waters linked to dinucleotides of a specific NtC/sequence combination were transferred to a reference dinucleotide. Finally, we computed water probability density distributions through Fourier averaging, separately for waters associated with the base and the sugar-phosphate atoms. Peaks in the hydration densities are referred to as Hydration Sites (HSs), and unveil the intricate interplay between base and sugar-phosphate hydration with respect to the sequence and conformation of DNA. The identified hydrated dinucleotide building blocks allowed us to subsequently calculate DNA hydration by determining the probability of water density distributions.

In this poster, we present an overview of our findings and discuss the potential applications of hydrated DNA building blocks for predicting DNA hydration. Our data and predictions are readily accessible for browsing and visualization at the website watlas.datmos.org/watna.

1. Biedermannová L. & Schneider B.: Structure of the ordered hydration of amino acids in proteins: analysis of crystal structures. Acta Crystallographica D71, 2192-2202 (2015).
2. Černý J., Schneider B. & Biedermannová L.: WatAA: Atlas of Protein Hydration. Exploring synergies between data mining and ab initio calculations. Phys. Chem. Chem. Phys. 19, 17094 (2017).
3. Biedermannová L. & Schneider B.: Hydration of proteins and nucleic acids: Advances in experiment and theory. A review. Biochimica et Biophysica Acta - General Subjects 1860, 1821-1835 (2016).
4. Schneider B. & Berman H.M.: Hydration of the DNA Bases Is Local. Biophysical J. 69, 2661-2669 (1995).
5. Schneider B., Patel K. & Berman H.M.: Hydration of the Phosphate Group in Double-Helical DNA. Biophysical J. 75, 2422-2434 (1998).
6. Biedermannová L., Černý, J., Malý, M., Nekardová, M., Schneider, B.: Prediction of DNA hydration from knowledge-based hydrated building blocks. Acta Cryst. D78, 1032-1045 (2022).

P5

CONFORMATIONAL STUDY OF SMALL CARBON RINGS IN LIGANDS

Gabriela Bučeková[1], Viktoriia Doshchenko[1], Radka Svobodová[1,2]

[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[2]CEITEC - Central European Institute of Technology, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
436961@mail.muni.cz

The primary repository not only for protein structures, Protein Data Bank (PDB), allows access to numerous structural data for biomacromolecules. However, due to the large amount of deposited structures may lead to the presence of errors in data. Early validation approaches primarily focused on the geometric properties of standard biomacromolecular residues, leading to the release of PDB validation reports. Later, this report extends to ligand validation, but some validation aspects are still not covered.

This study focuses on one of these overlooked areas of validation. We examined the conformations of basic rings found within small molecules in the PDB, including cyclopentane, cyclohexane and benzene. Inaccurate determination of ring conformation within ligands can significantly influence the geometric properties of a macromolecule or molecular fragments in the surrounding area. Our analysis extends to investigating the underlying reasons for selecting energetically unfavourable conformations of rings.

P6

PREDICTING THE EFFECTS OF MUTATIONS ON PROTEIN-PROTEIN INTERACTIONS VIA SELF-SUPERVISED MACHINE LEARNING

A. Bushuiev$^{\#1}$, R. Bushuiev$^{\#1,4}$, A. Filkin^{1}, P. Kouba^{1,2}, M. Gabrielova^{1}, J. Sedlar^{1}, T. Pluskal^{4}, J. Damborsky^{2,3}, S. Mazurenko^{2,3}, J. Sivic^{1}

[1]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague,
Jugoslavskych partyzanu 1580/3, 160 00 Prague, Czech Republic
[2]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk
University, Kamenice 5, 625 00 Brno, Czech Republic
[3]International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno,
Czech Republic
[4]Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nam. 542,
160 00 Prague, Czech Republic,
#Contributed equally
anton.bushuiev@cvut.cz

Fostering biomedical research and therapeutic advancement heavily relies on the design of protein-protein interactions (PPIs). However, machine learning methods for predicting the effects of mutations on protein-protein binding affinity suffer from systematic issues [1]. Most notably, existing methods exhibit poor generalization beyond training data due to reliance on small datasets of annotated mutations. Additionally, *in silico* scoring of binder variants with state-of-the-art tools is computationally expensive because of the need to simulate mutant structures. As a result, existing methods only enable the preselection of plausible single-point substitutions. However, fast and reliable methods that could combine the preselected substitutions to construct multi-point mutants with an enhanced binding affinity of a PPI are seriously lacking. These issues are particularly evident in our comprehensive case study of machine-learning-guided engineering of the staphylokinase protein for higher thrombolytic activity through the enhancement of its binding affinity towards plasmin [1].

As the unreliability of current methods stems from their dependence on small annotated data, in this project, we mine a vast amount of available unannotated PPIs in crystallized structures. Namely, we extract and cluster all protein-protein interactions from the entire Protein Data Bank. To achieve large-scale clustering, we develop a fast algorithm iDist for comparing pairs of protein-protein interfaces. The algorithm accurately approximates the well-established iAlign [2] method while being two orders of magnitude faster. Large-scale application of iDist reveals major, previously unaddressed issues with available datasets [3] of protein-protein interactions such as their high redundancy and low quality of train-test splits. Our data processing and analysis result in the construction of PPIRef – a novel clustered dataset of protein-protein interactions, superior in size and quality compared to existing alternatives [1].

Finally, we use the curated data to develop a novel self-supervised 3D equivariant deep learning model

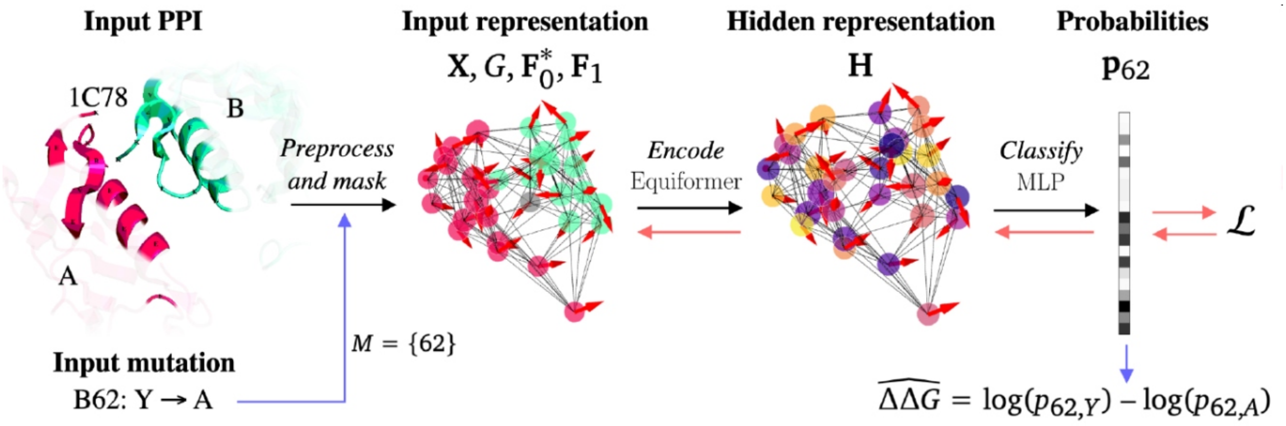


Figure 1. **Training and inference of PPIformer.** The black arrows depict the architecture of PPIformer, and the red arrows demonstrate the self-supervised training procedure for classifying masked amino acids. A single training step starts with randomly sampling a protein-protein interaction (in this example, the staphylokinase dimer A-B from PDB entry 1C78). After converting the interface into an oriented point-cloud representation, the features defining the side chains of randomly selected amino acids (e.g., 62 from chain B) are masked (shown by the grey circle). The model subsequently learns to classify the type of masked amino acids by acquiring an appropriate hidden representation of the whole interface. The blue arrows illustrate the zero-shot transfer of PPIformer to predicting binding G. In order to predict the mutational effect of substituting tyrosine (Y) with alanine (A) at position 62 in protein B, the corresponding amino acid is masked, and the probabilities are predicted with the trained model. Finally, the G is estimated based on the derived probabilities.

PPIformer [1]. The model leverages vast unannotated data by learning to solve a proxy task of predicting masked amino acids in the structures from our new PPIRef dataset. A more detailed overview of our approach is presented in Figure 1. We show that the proposed learning scheme enables PPIformer to capture the biochemical properties of amino acids and predict the experimental effects of mutations without any supervised training. This emergent property serves as a proof of concept for the approach, offering considerable hope for overcoming the data scarcity issue constraining existing methods for PPI design. Importantly, amino-acid-level representations of protein structures allow PPIformer to rapidly screen millions of protein variants without relying on costly molecular dynamics simulations of mutant structures.

Currently, we are focusing on unlocking the full potential of PPIformer. While it has demonstrated the ability to score mutational effects without any supervision, we expect it to become a powerful protein-design assistant as a result of further fine-tuning on the task of predicting binding G. Additionally, we are extending the method to other tasks, such as predicting binding energy or scoring docking poses. We expect the extensive multi-modal training of PPIformer to capture insights into complex biochemical phenomena such as epistasis or induced-fit mechanisms. Finally, we will utilize PPIformer in the next rounds of our computational staphylokinase design, which will be followed by thorough experimental validation. We expect the validated method to be broadly applicable in many biomolecular systems, including the design of antibodies and biopharmaceutics.

1. A. Bushuiev, Machine learning for the design of protein–protein interactions, Master's Thesis, Czech Technical University in Prague, Faculty of Information Technology, 2023.

2. M. Gao, J. Skolnick, iAlign: a method for the structural comparison of protein–protein interfaces, Bioinformatics, 26, (2010), 2259–2265.

3. R. Townshend, R. Bedi, P. Suriana, R. Dror, End-to-end learning on 3D protein structure for interface prediction, Advances in Neural Information Processing Systems,32, Curran Associates, Inc., 2019.

This work was supported by the Technology Agency of the Czech Republic through projects RETEMED and TEREP [TN02000122, TN02000122/001N] and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ [ID:90254] and ELIXIR [LM2023055].

P7

Mol*VS VISUALIZATION AND INTERPRETATION OF CELL IMAGING DATA ALONGSIDE MACROMOLECULAR STRUCTURE DATA AND BIOLOGICAL ANNOTATIONS

Aliaksei Chareshneu[1], Adam Midlik[1], Alessio Cantara[1], Crina-Maria Ionescu[1], Alexander Rose[2], Vladimír Horský[1], Radka Svobodová[1,3], Karel Berka[4], David Sehnal[1,3]

*[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University,
Brno, 602 00, Czech Republic
[2]Mol* Consortium - San Diego, CA, USA
[3]CEITEC - Central European Institute of Technology, Masaryk University, Brno, 625 00, Czech Republic
[4]Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, Olomouc, 779 00,
Czech Republic
alexcantara41@gmail.com*

Segmentation plays a crucial role in interpreting biological imaging data. As automated segmentation tools have advanced [1, 2], public repositories for imaging data have expanded their capabilities to include sharing and visualizing segmentations [3, 4]. This evolution has led to a growing demand for interactive web-based visualization of 3D volume segmentations. To address this challenge we have created Mol* Volumes and Segmentations (Mol*VS), which enables the interactive, web-based visualization of cellular imaging data, complemented by macromolecular data and biological annotations. Mol*VS seamlessly integrates with Mol* Viewer [5], a platform already adopted by several public repositories.

Mol*VS processes volumetric and segmentation data and deliver them to the Mol* Viewer VS extension, allowing the visualization of large datasets with low latency.

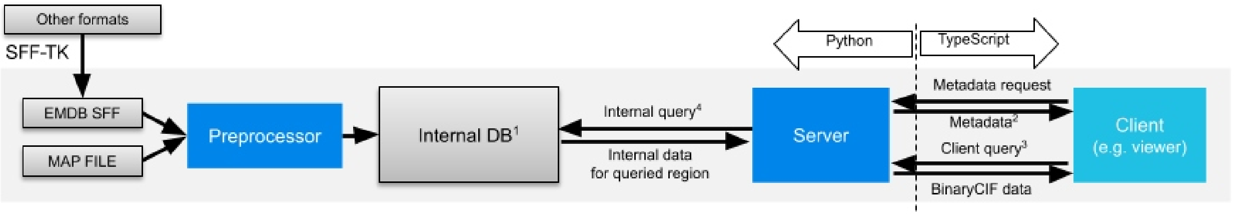
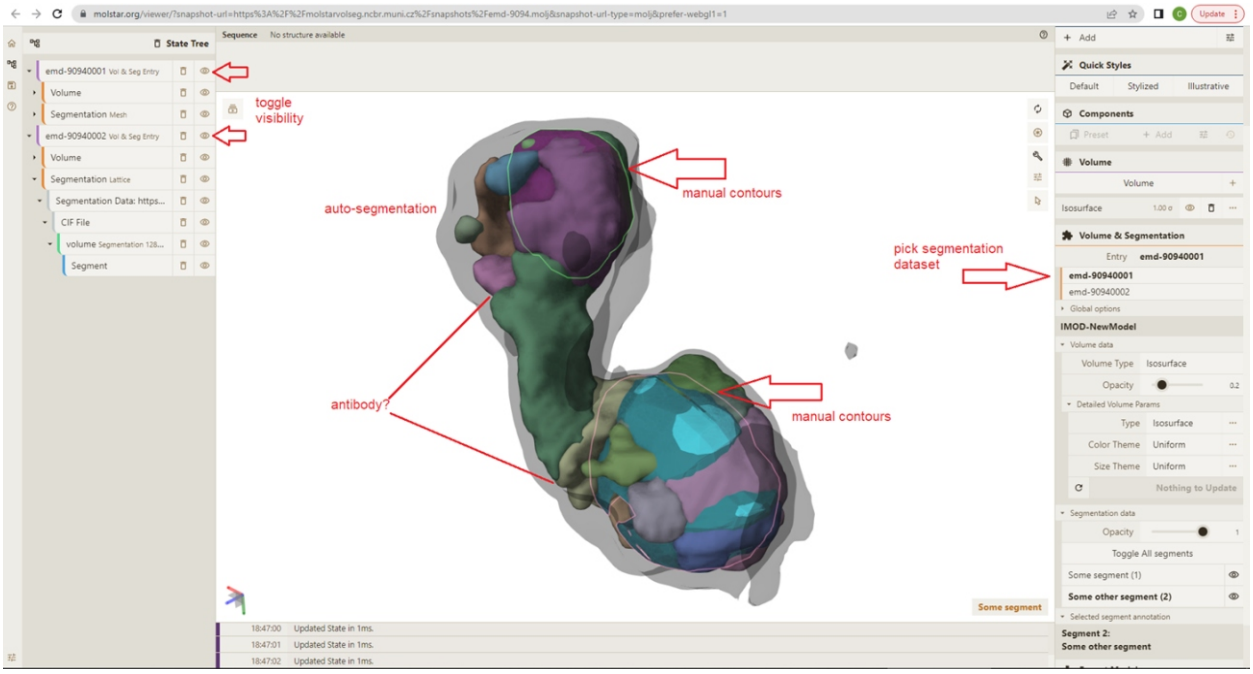


Figure 1. The four components composing Mol*VS and their functionalities



molstar.org/viewer/?snapshot-url=https%3A%2F%2Fmolstarvolseg.ncbr.muni.cz%2Fsnapshots%2Femd-9094.molj&snapshot-url-type=molj&prefer-webgl1=1

Sequence No structure available

Isosurface 1.00 σ

Entry emd-90940001

Volume Type Isosurface

Opacity 0.2

Type Isosurface

Color Theme Uniform

Size Theme Uniform

Opacity 1

Segment 2:
Some other segment

```
18:47:00 Updated State in 1ms.
18:47:01 Updated State in 1ms.
18:47:02 Updated State in 1ms.
```

Figure 2. An example of Mol*VS functionality. A comparison of different segmentation methods applied to a large Bacteriophage dataset correlated with 3D Volumes, Segmentation, Fitted Model and Model Annotation.

These functionalities are allowed by the presence of four major components: a pre-processor, an internal database, a server module, and a client module (**Fig. 1**).

An example of the Mol*VS functionalities is shown in **Figure 2**.

Mol*VS provides access to all EMDB and EMPIAR entries featuring segmentation datasets, supporting the visualization of data generated across a broad spectrum of electron and light microscopy experiments. Mol*VS is an open-source solution, freely accessible at https://molstarvolseg.ncbr.muni.cz/.

1. Kievits A.J., Lane R., Carroll E.C., Hoogenboom J.P. How innovations in methodology offer new prospects for volume electron microscopy. J Microsc. 2022; 287:114–137.

2. Thomas R.M., John J.Radhakrishnan B. A review on cell detection and segmentation in microscopic images. Proceedings of IEEE International Conference on Circuit, Power and Computing Technologies 2017 (ICCPCT 2017). 017; Kollam, IndiaInstitute of Electrical and Electronics Engineers (IEEE).

3. Iudin A., Korir P.K., Salavert-Torres J., Kleywegt G.J., Patwardhan A. EMPIAR: a public archive for raw electron microscopy image data. Nat. Methods. 2016; 13:387–388.

4. Iudin A., Korir P.K., Somasundharam S., Weyand S., Cattavitello C., Fonseca N., Salih O., Kleywegt G.J., Patwardhan A. EMPIAR: the Electron Microscopy Public Image Archive. Nucleic Acids Res. 2023; 51:D1503–D1511.

5. Sehnal D., Bittrich S., Deshpande M., Svobodová R., Berka K., Bazgier V., Velankar S., Burley S.K., Koča J., Rose A.S. Mol*Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. 2021; 49:W431–W437.

P8

APPLICATION AND ANALYSIS OF PROTEIN LANGUAGE MODELS FOR PREDICTING MEMBRANE INTERACTING PEPTIDE REGIONS

Máté Csepi[1], Gábor Erdős[2], Zsuzsa Dosztányi[2], Tamás Hegedűs[1,3]

[1]Department of Biophysics and Radiation Biology, Semmelweis University, 1085 Budapest, Hungary
[2]Department of Biochemistry, Eötvös Loránd University, 1117 Budapest, Hungary
[3]ELKH-SE Biophysical Virology Research Group, Eötvös Loránd Research Network,
1052 Budapest, Hungary
hegedus.tamas@hegelab.org

Interaction between proteins and lipids plays a crucial role in numerous cellular processes. Similar to protein-protein interactions, the involved peptide segments may be intrinsically disordered regions (IDRs) and their dynamics decreases upon lipid binding that may even include gaining secondary structures. We had collected proteins with lipid-interacting IDRs based on experimental data and named the interacting segments membrane molecular recognition features (MemMoRFs; https://memmorf.hegelab.org) [1].

Now we aimed to use this dataset to establish a predictor to overcome the tedious experiments for identification of membrane interacting IDRs. Since the available data set is small for conventional machine learning methods [2], we employed protein language models (pLMs), including T5 and Ankh. We used logistic regression to select important features of pLM embeddings. The analysis of these feature subsets indicates that MemMoRF properties are encoded in more features in Ankh than in T5 pLM embeddings. This phenomenon most likely contributes to better performance of Ankh in general when all features are used in predictions. However, we demonstrate that subsets of features can decrease the noise, thus increase the performance of per residue embedding-based neural networks for MemMoRF prediction. Our best performing model exhibits AUC, MCC, and F1 of 0.885, 0.655, and 0.827, respectively. We used this model to predict MemMoRFs in the human proteome, which predictions and also the predictor are available publicly at https://plmmorf.hegelab.org.

1. G. Csizmadia, G. Erdős, H. Tordai, R. Padányi, S. Tosatto, Z. Dosztányi, T. Hegedűs (2020). The MemMoRF database for recognizing disordered protein regions interacting with cellular membranes. *Nucleic Acids Research*, **49**, D355–D360. https://doi.org/10.1093/nar/gkaa954.

2. S. Basu, T. Hegedűs, L. Kurgan (2023). CoMemMoRFPred: sequence-based prediction of MemMoRFs by combining predictors of intrinsic disorder, MoRFs and disordered lipid-binding regions. *Journal of Molecular Biology*. https://doi.org/10.1016/j.jmb.2023.168272.

Support of grants NKFI-127961and NKFI-137610 to T.H. and National Academy of Scientist Education fellowships to M.C. and T.H. are acknowledged. We thank the colleagues of the Governmental Information-Technology Development Agency for computational resources on the Komondor GPU cluster.

P9

PREDICTING SARS-CoV-2 VARIANT FITNESS USING PROTEIN STRUCTURE INFORMATION

G. Cia[1,2], J. Kwasigroch[1,2], F. Pucci[1,2], M. Rooman[1,2]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles, Brussels, Belgium
[2]Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium
gabriel.cia@ulb.be

Genomic surveillance has played a fundamental role in the SARS-CoV-2 pandemic as it allowed the early identification, characterization and mitigation of dangerous viral strains. It would be greatly facilitated if accurate predictors of the impact of SARS-CoV-2 variants on virus fitness were available. However, most of the variant severity prediction models have limited predictive power as they only rely on SARS-CoV-2 sequence data, which prevents them from providing a molecular-level understanding of viral fitness and evolution.

To fill this gap, we developed SpikePro [1, 2], a fast and accurate structure-based model that predicts the fitness of SARS-CoV-2 variants from the sequence and structure of the spike protein located on the virus surface. Viral fitness indicates how efficiently a virus produces infectious progeny and is thus an indicator of its dangerousness. The model is based on the combined effect of the variants on the stability of the spike protein and on its binding affinity for the angiotensin converting enzyme 2 (ACE2) known to be the entry point of the virus into the cells, and for a set of neutralizing antibodies (nAbs). The predicted variant fitness is based on a combination of three components:

- Inter-human viral transmissibility, $^{S}_{i}$, computed from the change in folding free energy (G_S) be-

Table 1. Fitness values and R_0 values predicted by SpikePro, and epidemiological R_0 values for the major circulating SARS-CoV-2 lineages in the UK with respect to the Wuhan lineage. Epidemiological R_0 values were calculated by fitting a simple exponential growth model using GISAID data (see [2] for details). Values with an asterisk (*) were used to fit the model's parameters.

Variants	Transmissibility (log $^{S}_{i}$)	Infectivity (log $^{ACE}_{i}$)	Immune escape (log $^{nAb}_{i}$)	Global fitness (log $_{i}$)	R_0 epidemiological	R_0 predicted
Wuhan	0	0	0	0	2.2^*	2.2^*
Alpha	2.2	0.4	-0.2	2.3	6.1^*	6.1^*
Delta	1.6	0.3	1.0	2.8	8.1	7.0
Omicron	5.8	0.2	1.0	7.0	15.6	14.1

tween the spike protein variant *i* and the wild-type Wuhan strain. G_S is estimated using our in-house protein stability predictor *PoPMuSiC* [3].

- Viral infectivity, $^{ACE2}_{i}$, computed from the change in binding free energy (G_B) between the variant and wild-type Wuhan complex of the spike protein with ACE2, which we estimate using our in-house protein binding affinity predictor *BeAtMuSiC* [4].
- Viral immune escape, $^{nAb}_{i}$, computed from the average change in G_B between the variant and the wild-type spike protein in complex with a set of nAbs.

We predict the global fitness $_i$ of a variant as the product of these three fitness components (see [1] for technical details). We also predict the basic reproduction rate (R_0), a commonly monitored indicator for new emerging variants that describes the average number of secondary infections that each infected person causes in a naive population. The decomposition of the viral fitness into these three components representing viral transmissibility, infectivity and immune escape helps users understand which molecular mechanism is driving the overall fitness evolution of the SARS-CoV-2 variants with respect to the original Wuhan lineage.

We used SpikePro to predict and analyze SARS-CoV-2 variants that had high circulation levels worldwide during the pandemic. The predicted transmissibility, infectivity, immune escape and global fitness of the major circulating SARS-CoV-2 lineages (Alpha, Delta and Omicron) with respect to Wuhan are reported in Tab. 1. We observe that SpikePro correctly ranks the lineages in terms of global fitness: (Wuhan) < (Alpha) < (Delta) < (Omicron). Moreover, we correctly predict that Omicron has a much higher transmissibility but is less infectious than other lineages, which is in agreement with empirical data [5]. Furthermore, as shown in Tab. 1, the predicted R_0 value for the Delta and Omicron variants, which were not used for parameter fitting, agree very well with the epidemiological data, with an error of the order of 10%.

For the most recently circulating Omicron subvariants X.BB.1.5 and EG.5.1, SpikePro predicts R_0 and global fitness values that are similar to the original Omicron variant. In particular, it predicts a decrease in transmissibility and infectivity but a significant increase in immune escape, all of which is in very good agreement with epidemiological and clinical data which show that these subvariants result in milder clinical symptoms and have high immune escape [6].

In our paper [ref], we also present the results of a retrospective study to show how the webserver could have predicted the appearance of certain lineages that improve fitness, spread fast and escape human immune defenses. We calculated the average R_0, infectivity, transmissibility, and ability to escape from the immune system over all sequences in the GISAID database, and found that the predictions made by SpikePro agree remarkably well with epidemiological and experimental data. Note that this retrospective study does not include any parameter fitting, which greatly adds confidence in our tool.

We developed a user-friendly webserver (http://babylone.3bio.ulb.ac.be/SpikePro/) with which researchers without any bioinformatic background can easily run SpikePro. Using the variant spike protein sequence as input, the webserver provides the overall fitness value of the variant, its fitness components and its basic reproduction rate R_0. It also integrates external experimental and epidemiological data and provides a 3D visualization of the mutated spike protein structure.

1. F. Pucci and M. Rooman, *Viruses*, **13**, 935 (2021).

2. G. Cia, J. M. Kwasigroch, M. Rooman, F. Pucci, *Bioinformatics*, **38**, 4418 (2022).

3. Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, M. Rooman, *Bioinformatics*, **25**, 2537 (2009).

4. Y. Dehouck, J. M. Kwasigroch, M. Rooman, D. Gilis, *Nucleic Acids Research*, **41**, W333 (2013).

5. K. P. Y. Hui, J. C. W. Ho, M. Cheung, K. Ng, R. H. H. Ching, K. Lai, T. T. Kam, H. Gu, K.-Y. Sit, M. K. Y. Hsin, T. W. K. Au, L. L. M. Poon, M. Peiris, J. M. Nicholls, M. C. W. Chan, *Nature*, **603**, 715 (2022).

6 R. Uraki, M. Ito, Y. Furusawa, S. Yamayoshi, K. Iwatsuki-Horimoto, E. Adachi, M. Saito, M. Koga, T. Tsutsumi, S. Yamamoto, A. Otani, M. Kiso, Y. Sakai-Tagawa, H. Ueki, H. Yotsuyanagi, M. Imai, Y. Kawaoka, *The Lancet Infectious Diseases*, **23**, 30 (2023).

P10

EXTENSION OF THE SUGRES COARSE-GRAINED MODEL OF POLYSACCHARIDES TO HEPARIN

A. Danielsson, S.A. Samsonov, A. Liwo, A.K. Sieradzan

Faculty of Chemistry, University of Gdańsk, ul. Wita Stwosza 63, 80-308 Gdańsk, Poland
a.danielsson.317@studms.ug.edu.pl

The glycosaminoglycan heparin (HP) is an unbranched periodic polysaccharide composed of negatively charged disaccharide units and involved in key biological processes, including anticoagulation, angiogenesis, and inflammation. The considerable size and flexibility of naturally occurring HP as well as the predominantly electrostatic nature of its interaction with proteins renders it a particularly difficult target in all-atom molecular dynamics (MD) simulations of the molecular mechanisms underlying the biologically relevant multiscale processes. Therefore, application of coarse-grained approaches is potentially promising to model HP-containing molecular systems.

We have extended the coarse-grained SUGRES-1P model (Fig. 1) of polysaccharides [1] to HP and modified the interaction energy function to account for a shift of the interaction centres and to enable a direct modification of the electrostatic energy term weight. The implemented parameters were previously obtained using all-atom MD simulations [1,2] with the GLYCAM06 force field [3]. With this modification, we were able to apply the SUGRES-1P force field in microsecond-long MD simulations of free HP oligosaccharides ranging from degree of polymerization 6 to 68. The modelled HP chains exhibited remarkable similarity to experimentally determined HP molecules [4,5] in terms of their global structural characteristics. A comprehensive analysis of the constituent energy term weights and ion concentration, represented by the Debye-Hückel parameter , indicates that long HP chains are characterized by coiled conformations governed predominantly by electrostatic interactions established between the charged residues.

We integrated the SUGRES-1P model into the coarse-grained UNICORN model [6,7], enabling microsecond-scale MD simulations of HP interactions with single- and multi-domain proteins. This achievement represents a significant milestone, as it is the first time a "bottom-up" physics-based approach has been used for coarse-grained modelling of HP chains, while maintaining compatibility with other biomolecule classes within the UNICORN modelling package.

1. E. Lubecka, A. Liwo, *J. Chem. Phys.*, **147**, (2017), 115101.
2. S. A. Samsonov, E. A. Lubecka. K. K. Bojarski, R. Ganzynkowicz, A. Liwo, *Biopol.* ***110***, (2010), e23269.
3. K. N. Kirschner, A. B. Yongye. S. M. Tschampel, J. Gonzalez-Outeirino, C. R. Daniels. B. L. Foley, R. J. Woods, *J. Comp. Chem.* ***29***, (2008), 622-655.
4. S. Khan, J. Gor, B. Mulloy, S. J. Perkins, *J. Mol. Biol.* ***395***, (2010), 504-521.
5. G. Pavlov, S. Finet, K. Tatarenko, E. Korneeva, C. Ebel, *Eur. Biophys.* ***32***, (2003), 437-449.
6. A. Danielsson, S. A. Samsonov, A. Liwo, A. K. Sieradzan, *J. Chem. Theory Comput.* ***19***, (2023), 6023-6036.
7. A. Liwo, C. Czaplewski, A. K. Sieradzan, et al., *Prog. Mol. Biol. Transl. Sci.* ***170***, (2020), 73-122.

This work was supported by the Polish National Science Centre, grants UMO-2018/31/G/ST4/00246, UMO-2021/40/Q/ST4/00035 and UMO-2017/27/B/ST4/00926. Calculations were conducted using the cluster at the Faculty of Chemistry, University of Gdańsk.

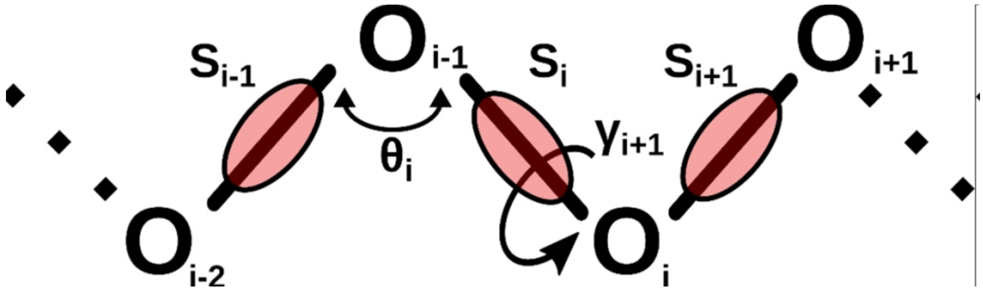


Figure 1. The interaction sites are united sugar rings, represented by transparent red ellipsoids, located half-way between glycosidic oxygen atoms (white spheres) which are not interaction sites but serve to define the geometry of the polysaccharide molecule. The virtual bonds connecting the oxygen atoms are shown as thick black lines. The geometry of the polysaccharide chain is defined by the virtual bond angles $_i$ and torsional angles $_i$.

P11

MAPPING AND CHARACTERIZATION OF THE HUMAN MISSENSE VARIATION UNIVERSE USING ALPHAFOLD 3D MODELS

E.I. Timothy, S.A. Islam, G. Hanna, M.J.E. Sternberg, A. David

Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Imperial College London, London SW7 2AZ, UK
aleesia.david09@imperial.ac.uk

In the last few years, we have witnessed major developments in both protein structure prediction, as exemplified by AlphaFold 3D models [1], and, most recently, in language-based models for the prediction of genetic variants, including the recently released AlphaMissense [2]. However, these predictors do not provide insights into the structural basis of the phenotypic effect.

Three-dimensional protein structures allow us to perform an atom-based analysis of the consequence of an amino acid substitution and models generated with the deep learning algorithm, AlphaFold, provide a unique opportunity for atom-based analysis of human missense variants. Our group has developed the Missense3D portal to provide structure-based interpretation of missense variants [3]. Here we present the results obtained with a new pipeline that we have developed to automatically identify accurately modelled amino acid regions that can be used for variant characterization with our in-house Missense3D algorithm.

The recommended AlphaFold pLDDT threshold for an accurately modelled residue is 70. When using this threshold for the query residue, the accuracy of the atom-based predictions calculated using Missense3D on AlphaFold models was 0.66, MCC 0.36, TPR/FPR 5.1. We then compared these results to those obtained with lower pLDDT scores, and after introduction of the PAE matrix score, on a benchmark dataset of 10,085 human proteins harbouring 84,827 missense variants.

We show that, when the model accuracy of the environment surrounding the query residue (E-plDDT-5Å) is considered, an E-plDDT-5Å 60 provides similar accuracy, MCC and TPR/FPR to that obtained using the plDDT threshold 70 for the query residue alone but increases the number of residues for which an atom-based analysis can be performed.

We applied this new E-plDDT-5Å 60 threshold to a total of 8,965,659 residues corresponding to 16,325 reviewed human UniProt sequences of lengths 2,700 amino acids. When using this threshold, 6,169,173 human residues (68.8% of the proteome) are modelled with sufficient quality to allow an atom-based analysis of the query residue and its surrounding environment. At the variant level, confident predictions could be obtained for 4,405,910 (65.9%) out of 6,700,719 unique missense variants mined from the UniProt homo_sapiens_variation.txt database.

In conclusion, AlphaFold 3D models offer a unique opportunity to understand the consequences of amino acid substitutions on protein structure, thus complementing existing evolutionary-based methods.

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al.: Highly accurate protein structure prediction with AlphaFold. Nature, 596, (2021), 583-589.
2. Cheng J, Novati G, Pan J, Bycroft C, Žemgulyt? A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, et al.: Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (2023),7492.
3. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE: Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J. Mol. Biol., 431, (2019), 2197-2212.

P12

COMPUTATIONAL RESOURCES FOR ANALYZING TRANSMEMBRANE PROTEIN STRUCTURES, AND TOPOLOGY

Laszlo Dobson[1,2], **Gabor Tusnady**[1]

[1]Research Centre for Natural Sciences, Magyar Tudosok Korutja, 1117-Budapest, Hungary
[2]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117-Heidelberg, Germany
dobson.laszlo@ttk.hu

The Membrane Protein Bioinformatics Research Group hosts and maintains several widely used resources related to transmembrane protein structures.

In 2021, AlphaFold2 (AF2) opened new frontiers for almost all fields of structural biology and provided 3D structures for almost all known protein sequences. In the Transmembrane AlphaFold database (TmAlphaFold database, https://tmalphafold.ttk.hu/) we use a simple geometry-based method to visualize the likeliest position of the membrane plane using AF2 structures as a source. In addition, we calculate several parameters to evaluate the location of the protein into the membrane. This also allows the TmAlphaFold database to show whether the predicted 3D structure is realistic or not.

We also overhauled several other popular resources and combined them in the The UNIfied database of TransMembrane Proteins (UniTmp, https://www.unitmp. org/). UniTmp is a comprehensive and freely accessible resource of transmembrane protein structural information at different levels, from localization of protein segments, through the topology of the protein to the membrane-embedded 3D structure. We not only annotated tens of thousands of new structures and experiments, but we also developed a new system that can serve these resources in parallel. UniTmp is a unified platform that merges TOPDB (Topology Data Bank of Transmembrane Proteins), TOPDOM (database of conservatively located domains and motifs in proteins), PDBTM (Protein Data Bank of Transmembrane Proteins), and HTP (Human Transmembrane Proteome) databases and provides interoperability between them.

In the near future we plan to integrate more databases and web servers into the framework of UniTmp, so researchers will be able to find all membrane protein related information at one place.

P13

ON MODELLING SIGNALLING AMYLOID MOTIFS

W. Dyrka[1], **J. Gałązka**[1], **M. Gąsior-Głogowska**[1], **K. Pysz**[1], **M. Szefczyk**[2], **N. Szulc**[1,3]

[1]Katedra Inżynierii Biomedycznej, Wydział Podstawowych Problemów Techniki, Politechnika Wrocławska, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[2]Katedra Chemii Bioorganicznej, Wydział Chemiczny, Politechnika Wrocławska, Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
[3]Katedra Fizyki i Biofizyki, Wydzial Biotechnologii i Nauk o Zywnosci, Uniwersytet Przyrodniczy we Wroclawiu, Norwida 25, 50-375 Wroclaw, Poland
witold.dyrka@pwr.edu.pl

Amyloid motifs are very short domains (typically around 25 amino acids) that facilitate protein aggregation into a polymeric fibrillary structure or the amyloid fold. Apart from the involvement in pathological amyloidoses, the amyloid fold is vital for many physiological functions including oligomerization and signal transduction. Despite considerable sequential diversity and partial similarity to intrinsically disordered regions, known functional amyloid motifs typically assume the beta-arch conformation and aggregate into beta-solenoids through stacking of the beta strands. Identification of functional amyloids in sequence databases is difficult. Existing computational tools for assessing propensity to form amyloids typically focus on the amyloidogenic hotspots or short peptides forming amyloid fibrils *in vitro*. These tools are not calibrated for meaningful searches in entire genomes as it is estimated that most proteins contain amyloidogenic hotspots [1]. An alternative approach is to model particular families of amyloidogenic sequences, which can be a viable option for already identified motifs (e.g. Pfam profiles HET-s_218-289 and RHIM). However, as amyloidogenic motifs are relatively short and quite diverse internally, traditional tools based on k-mers or profile Hidden Markov Models do not offer enough statistical power for more generalised genome-wide searches.

Signalling amyloid motifs work in pairs (or triplets), where one motif triggers the other to assume the amyloid fold in a prion-like manner. To date, the most successful searches relied on identification of larger domains associated with the amyloid motifs combined with filtering based on genomic proximity and sequential similarity of potentially cooperating sequences. This led to identification of around thousand signalling amyloids representing more than a dozen of families associated with the Nod-Like Re-

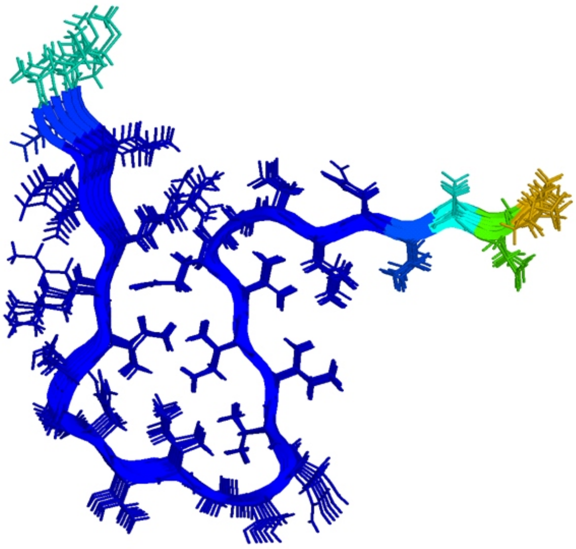


Figure 1. Model of HRAM4 homooligomer based on a custom MSA with 111 sequences created using the template-free ColabFold advanced notebook [7]. Residues colored according to lDDT.

ceptors (or NLR, innate immune system proteins) in filamentous fungi, bacteria and archaea [2, 3]. These datasets have been used to develop machine learning models capable of accounting for non-local dependencies resulting from the spatial fold. Our generalised model of several families of signalling amyloid motifs, based on the probabilistic context-free grammars (git.e-science.pl/wdyrka/pcfg-cm), demonstrated good specificity and sensitivity enabling searches in collections of thousands NLR-associated proteins [1]. This facilitated discovering new motif families in sac fungi and made possible exploring the signalling amyloidosome of filamentous basidiomycota [4]. Currently, we develop deep learning methods aimed at achieving higher specificity enabling genome-wide searches without the need for prefiltering (github.com/jakub-galazka/asmscan-lstm, github.com/chrispysz/amylotool-console).

In some cases even relatively small high-quality alignments of several dozens of sequences provide enough information to predict structural models of amyloid structures made of several copies of a motif. For example, using AlphaFold 2, we obtained a plausible model of HET-s related motif [2, 4] (Fig. 1). However, the main issue with applying the general protein structure prediction protocols to amyloidogenic peptides is the method indifference to single point mutations affecting the formal charges, even though the feature was shown experimentally to alter the aggregation propensity of amyloids [5, 6]. Despite this limitation, AlphaFold-style predictions of a potential fold of the motif can be a useful step preceding the more fine-tuned investigation with other methods, such as molecular dynamics [6].

The short length of typical amyloidogenic peptides means that they can be synthesised relatively easily, although the process has its peculiarities due to high aggregation [1, 4-6]. The use of synthetic peptides enables quite rapid experimental verification of the modelling *in vitro*. For example, infrared spectroscopy can be used to establish the structural content and rigidity reflecting the aggregation stage. often in conjunction with an imaging method such as atomic force microscopy. In addition the kinetics of aggregation is typically evaluated through the Thioflavin T assay [4]. As properties of functional amyloids are often very sensitive to even slight changes in environmental conditions such as temperature, pH or presence of certain ions, we postulate the *in vitro* validation of the aggregation process to be performed over the grid of conditions.

The ultimate goal is predicting and modelling their heterotypic interactions, such as cross-seeding and hetero-aggregation, with the hope of deciphering potential triggers of amyloidosis [8]. However, as of now, data on cross-interactions of pathological amyloids are too scarce and sparse for training robust machine learning models. One possibility is to retreat towards threading-based methods [9]. At the same time, the growing number of evolutionary coupled pairs of signalling amyloid motifs opens perspectives also for the machine learning approaches. So far, the identification of NLR-related interacting pairs relied on genomic proximity and sequential similarity of candidate motifs. However, in some cases at least one of these conditions could be unmet [4], which necessitates a direct prediction of the interaction. Experimental evidence for the presence of so-called gate-keeper residues of the aggregation process [5, 10] suggest that such prediction is a viable option at least for functional amyloids. It can be expected that concerted use of modelling and experimental methods could pave a way for decoding interaction networks of signalling amyloids within and between genomes, better understanding their interactions with other proteins (e.g. beta-solenoid tandem repeats), and even enable their use for biocomputing.

1. W. Dyrka, M. Gąsior-Głogowska, M. Szefczyk, N. Szulc, BMC Bioinformatics, 22, (2021), 222.
2. A. Daskalov, W. Dyrka, S. Saupe, *Sci Rep*, **5**, (2015), 12494.
3. W. Dyrka, V. Coustou, A. Daskalov, A. Lends, et al., *J Mol Biol*, **432**, (2020), 6005.
4. J.W. Wojciechowski, E. Tekoglu, M. Gąsior-Głogowska, V. Coustou, et al., *PLoS Comput Biol*, **18**, (2022), e1010787.
5. N. Szulc, M. Gąsior-Głogowska, J.W. Wojciechowski, M. Szefczyk, et al., *Int J Mol Sci*, **22**, (2021), 5127.
6. N. Szulc, M.E. Gąsior-Głogowska, P. Żyłka, M. Szefczyk, et al., ssrn.com/abstract=4521809.
7. J. Jumper, R. Evans, A. Pritzel, T. Green, et al., *Nature*, **596**, (2021), 583.
8. R.P. Friedland, M.R. Chapman, *PLoS Pathog,* **13**, (2017), e1006654.
9. J.W. Wojciechowski, W. Szczurek, N. Szulc, M. Szefczyk, et al., *bioRxiv*, 2022.07.07.499150.
10. A. Daskalov, D. Martinez, V. Coustou, N. El Mammeri, et al., *Proc Natl Acad Sci USA*, **118**, (2020), e2014085118.

P14

AHOJ DB: A PDB-WIDE ASSIGNMENT OF APO & HOLO RELATIONSHIPS BASED ON INDIVIDUAL PROTEIN-LIGAND INTERACTIONS

Christos Feidakis[1], Radoslav Krivak[2], David Hoksza[2], Marian Novotny[1]

[1]Charles University, Faculty of Science, Department of Cell Biology, Czech Republic
[2]Charles University, Faculty of Science, Department of Software Engineering, Czech Republic
christos.feidakis@natur.cuni.cz

A repertoire of scenarios in structural biology requires access to multiple snapshots of a protein. From studying protein dynamics to unveiling cryptic binding sites, from assessing the effectiveness of ligand binding site prediction software to building datasets for training such machine learning predictors, a single protein structure is rarely sufficient to capture or explain the variability of a protein.

The availability of both bound (holo) and unbound (apo) forms of a protein structure is essential for making meaningful comparisons and drawing robust conclusions. The few existing resources that provide access to such data are limited either in terms of protein coverage or in the number of structure pairs provided, which does not always reflect the conformational variance represented by the structures deposited in the Protein Data Bank (PDB).

Here, we use a previously designed application (AHoJ, Apo-Holo Juxtaposition) to perform an extensive search for apo-holo pairs for each individual protein-ligand interaction across the PDB (excluding interactions with peptides and nucleic acids). We assemble the results of ~500,000 small molecule interactions into a database that can be used to train and evaluate predictors, discover potentially druggable proteins, and reveal associations that can confirm existing hypotheses or expose protein- and ligand-specific relationships like order-to-disorder transitions that were previously obscured by intermittent or partial data, or discover specific binding properties of individual ligands.

P15

TOWARD A DATABASE OF GENETICALLY ENCODED NON-CANONICAL AMINO ACIDS

Carolina F. Rodriges[1,2], Antoine Daina[3,4], Marta A.S. Perez[3,4], Vincent Zoete[3,4], Bohdan Schneider[1], Gustavo Fuertes[1]

[1]Institute of Biotechnology of the Czech Academy of Sciences, Czech Republic
[2]Instituto Politécnico de Setúbal, Escola Superior de Tecnologia de Barreiro, Portugal
[3]SIB Swiss Institute of Bioinformatics Lausanne, Switzerland
[4]University of Lausanne, Switzerland
gustavo.fuertes@ibt.cas.cz

SwissSideChain [1] is the only database specifically devoted to non-natural sidechains and displays general properties (physical, structural and molecular data) for 210 amino acids in both L- and D-configuration. Additionally, the provided modeling tools permit the straightforward insertion of these unnatural residues into proteins in silico. However, it is not directly obvious what are the specific properties that make these unnatural amino acids useful or how to prepare proteins made of unnatural building blocks. Among the available methods, genetic code expansion (GCE) enables the ribosome-mediated introduction of a large number of non-canonical amino acids (ncAA) into polypeptides at virtually any target positions. GCE is based on codon reassignment via orthogonal pairs composed of an engineered aminoacyl-tRNA synthetase (aaRS) and its cognate tRNA. Thus, we hereby present our initial efforts to update SwissSideChain with new genetically encoded ncAA in order to make the resource more practically useful. By including key data, potential applications, and sequence information on aaRS/tRNA, we expect to guide researchers in obtaining tailor-made proteins-of-interest carrying ncAA.

1. David Gfeller, Olivier Michielin, Vincent Zoete, SwissSidechain: a molecular and structural database of non-natural sidechains, *Nucleic Acids Research*, Volume 41, Issue D1, 1 January 2013, Pages D327–D332, https://doi.org/10.1093/nar/gks991.

This research was funded by the Czech Academy of Sciences, grant RVO 86652036.

P16

VALINOMYCIN INTERACTIONS WITH WATER

J. Hašek [1], M. Dušek [2], I. Císařová[3], J. Dybal[4],T. Skálová[1], J. Dušková[1], J. Dohnálek[1]

[1]Institute of Biotechnology, Academy of Sciences, Průmyslová 595, Vestec, Czech Republic
[2]Institute of Physics, Academy of Sciences, Cukrovarnická 2, Praha, Czech Republic
[3]Faculty of Science, Charles University, Hlavova 2030, Praha, Czech Republic
[4]Institute of Macromol.Chemistry, Academy of Sciences, Heyrovského nám.2, Praha, Czech Republic
hasekjh@seznam.cz

Valinomycin is well known compound important for high selectivity of the transport of K^+ ions through lipophilic membranes. This study indicates possible formation of valinomycin tunnels through the membranes under specific conditions. Originally, the crystal structures of valinomycin were studied on crystals grown from strictly dehydrated solutions. This is probably the reason of low diffraction quality of crystals and unusually high *R* factors (R > 17 %). Eight structures in period 1975–1980 were not accepted into the CSD and thus their coordinates are lost. Water is undoubtedly important to satisfy the hydrophilic inner part of the valinomycin surface. Thus, we decided to uncover the role of water in the valinomycin structure and function.

Quantum chemical calculations

The quantum chemical calculations (QCC) were carried out using the density functional theory (DFT) with the B3LYP functional and the 6-31G(d) basis set employing the Gaussian 03 program package. Reliable indicator of the trapped water ionization is the distance between the water oxygens (neutral 3.2 Å, ionized 2.4 Å) (Figure 1).

Experimental

We prepared crystals and experimentally determined structures of:

1. Valinomycin complexed with two uncharged molecules water (*R*=3.9 %),

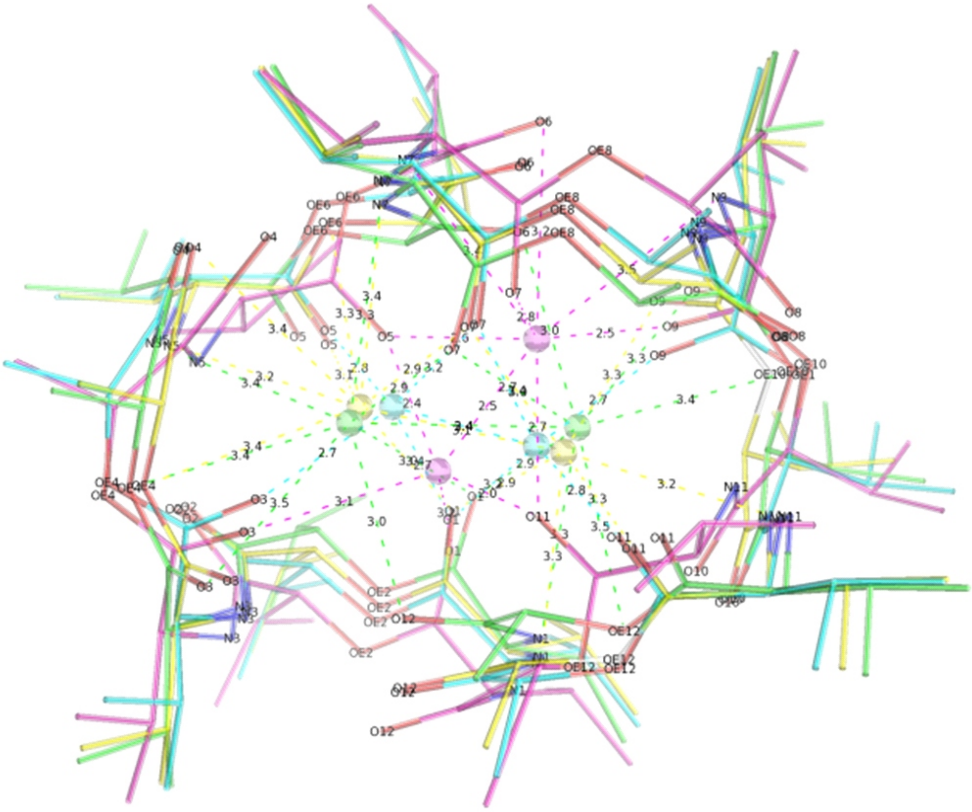


Figure 1. Superposition of valinomycin structures:

1. *valinomycin complexed with two neutral waters (diffraction experiment) yellow carbons and waters*
2. *valinomycin complexed with two neutral waters (calculated) green carbons and waters,*
3. *valinomycin complexed with hydronium $[H_5O_2]^+$ (calculated) blue carbons and waters,*
4. *valinomycin complexed with $[H_6O_2]^{2+}$ (calculated) magenta carbons and waters.*

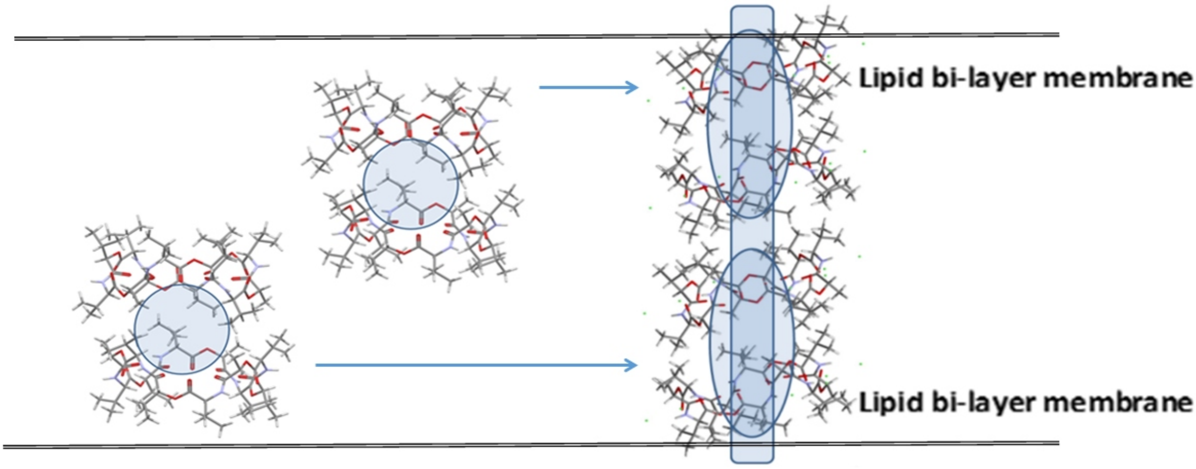


Figure 2. Two valinomycin dimers can form a water tunnel through the lipid bi-layer membrane. This configuration in membrane may be stabilized by chlorine ions similarly as the tunnels observed in the crystal structure.

2. Valinomycin dimers encapsulating >12 water molecules. The dimers are stacked in crystal to form infinite hydrophilic tunnels (higher R=7.9 % corresponds to mobility of waters in the inner tunnel).

Self-organization of valinomycin to form hydrophilic tunnels in hydrophobic membranes

The crystal structure is formed by large spherical valinomycin dimers encapsulating 12-15 water molecules (blue spheres in Fig. 2) and forming a disks with external diameter ~16 Å and the height about 20 Å (on the left side of Fig. 2). The disks are stacked in the crystal to form the infinite water filled tunnels (on the right side of Fig. 2), stabilized by hexagons of chlorine ions encapsulated in hydrophobic pockets. Interior of the tunnel is hydrophilic and the external surface of the tunnel is hydrophobic.

Conclusion

The quantum chemical calculations and the experiments confirmed that the positive charge of water molecules has small effect on the valinomycin conformation. The only reliable indicator of water ionization is the distance between water oxygens.

The main driving force for stabilization of the valinomycin tunnel are hexagons of chaotropic anions trapped in the hydrophobic pockets of external hydrophobic groups of the neighbor valinomycin molecules. The experimentally confirmed structure indicates possibility of formation of self-assembled hydrophilic tunnels through the hydrophobic bi-layers under special conditions (e.g. presence of small chaotropic anions).

The work was supported by the Czech Academy of Science project no. 86652036.

P17

ALPHAMISSENSE PERFORMANCE ON TRANSMEMBRANE PROTEINS

H. Tordai[1], T. Hegedűs[1,2]

[1]Dep. of Biophysics and Radiation Biology, Semmelweis Univ., Tűzoltó u. 37-47, Budapest 1094, Hungary
[2] HUN-REN-SU Biophysical Virology Research Group, Piarista u. 4, Budapest 1052, Hungary
hegedus.tamas@hegelab.org

Single amino acid substitutions can impact protein folding, dynamics, and function, often leading to severe pathological consequences. Distinguishing between benign and pathogenic substitutions is crucial for guiding research and therapeutic interventions. Unfortunately, the experimental investigation of these variants remains limited due to resource constraints. In response to this challenge, AlphaMissense [1] has emerged as a promising tool for predicting the pathogenicity of single nucleotide polymorphism (SNP) variants, surpassing other existing predictors.

Since transmembrane proteins are challenging for both experimental and computational approaches, we evaluated the performance of AlphaMissense on transmembrane proteins, utilizing ClinVar data for validation (positive: missense variants likely pathogenic and pathogenic; negative: likely benign and benign; reviewed: with at least one star). To ensure the reliability of our dataset, we exclusively considered TM topology predictions with a high confidence level (reliability score > 85) from the Human Transembrane Proteome [2]. Consequently, our dataset comprises 1,653 transmembrane proteins, with 1,228 and 6,370 variations located within TM and non-TM regions, respectively. Our evaluation reveals that AlphaMissense demonstrates remarkable performance, achieving F1 and MCC scores of 0.90 and 0.74 for TM regions and 0.82 and 0.69 for non-TM locations. This suggests that AlphaMissense is equally effective at predicting the pathogenicity of variants in both transmembrane and soluble regions of proteins.

Furthermore, we investigated variant predictions for selected TM ABC proteins associated with diseases, such as cystic fibrosis, to assess the distribution of pathogenic mutation in structural context. Moreover, recognizing that the accessibility of AlphaMissense predictions may be limited for many researchers and clinicians, we have developed a user-friendly, protein-centric web database. This resource, available at https://alphamissense.hegelab.org, offers an easy access to AlphaMissense data, enhancing the utility of this valuable tool.

Our study on AlphaMissense's performance in the context of membrane proteins, coupled with this web resource, promises to facilitate the broader utilization of AlphaMissense for research and clinical applications.

1. J. Cheng et al., *Science,* **381**, (2023), eadg7492.
2. G. E. Tusnády, Z. Dosztányi, I. Simon, *Nucleic acids research*, **33**, D275–D278.

This work has been supported by the National Research, Development and Innovation Office (grant number: K 137610).

P18

LARGE-SCALE APPLICATION OF PROTEINMPNN TO DESIGN CONFORMATION-SPECIFIC GPCR AGONISTS

Hannes Junker, Clara T. Schoeder

Institute for Drug Discovery, Leipzig University Medical Faculty, Leipzig, Germany

With more than 800 members G protein-coupled receptors (GPCRs) are the largest family of transmembrane receptors. Many GPCRs occupy variations of active states, each with a conformation-dependent signaling profile, a phenomenon termed biased signaling. Modulating the propensity for the occupation of certain states with fine-tuned agonists therefore holds great pharmacological promise. However, computational design of GPCR targeting ligands usually occurs on experimentally solved structures which represent only one snapshot within a conformational ensemble. Here we present our efforts to computationally design peptide agonists that can target conformational substates on the example of the Class A GPCR Growth Hormone Secretagogue Receptor (GHSR). To this end, we investigate the novel deep-learning design method ProteinMPNN which predicts an alternative amino acid sequence based on a provided backbone conformation. More precisely, we hypothesize that ProteinMPNN can be used to design peptide agonists, that are optimized for a specific receptor conformation. To provide coherent backbone variations, we utilized continuous frames from a ~36 s long MD simulation of the GHSR in complex with its endogenous peptide agonist Ghrelin and applied ProteinMPNN on the agonist. We demonstrate that ProteinMPNN is highly sensitive towards minor changes in the orthosteric binding pocket conformation by responding with a suitable agonist sequence. Given an adequate set of alternative receptor structures this approach provides the possibility to rapidly identify peptide sequences that address specific conformational states which in turn are associated with a specific signaling profile.

P19

DETERMINING STRUCTURAL CHARACTERISTICS OF NATIVE AND DE NOVO PROTEINS FOR IMPROVED PROTEIN DESIGN ALGORITHMS

Johannes A. Klier[1], Jens Meiler[1,2], Clara T. Schoeder[1]

[1]Institute for Drug Discovery, Leipzig University Medical Faculty, Leipzig, Germany
[2]Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, United States
klier@izbi.uni-leipzig.de

De novo proteins are derived from methods such as computational protein design in contrast to naturally evolved proteins. By definition, amino acid sequences of de novo proteins are not found in nature and are unmatched by natural sequences. De novo proteins designed with the software framework Rosetta often display characteristic biophysical behaviours like high thermostability and rigid structures, which limits protein flexibility. In order to design more natural proteins in the future, computational algorithms have to be improved. To achieve this, it is necessary to understand differences between natural and de novo proteins. For this purpose, a computational scoring system was developed, derived from structural and biophysical characteristics and predictions made by the deep residual network trRosetta. Various interresidue distance metrics showed clear differences between natural and de novo proteins. Predictions by trRosetta for pairwise C   distances between two residues were more accurate for de novo proteins than for natural proteins while predictions for interresidue orientations were worse. It was also found that de novo proteins contain more interresidue interactions than natural proteins and a lower fraction of noninteracting residues. A correlation between low absolute contact order and designed proteins could be established. These findings will help inform further protein design algorithm optimization to improve design of proteins with natural flexibility and function.

P20

ASSESSING THE EFFECTS OF MUTATIONS AND CORRECTOR MOLECULES ON APOE DYNAMICS VIA COMPARATIVE MARKOV STATE ANALYSIS

Jakub Kopko[1], Petr Kouba[1,2], Sérgio M. Marques[2,3], Joan Planas-Iglesias[2,3], Jiří Damborský[2,3], Stanislav Mazurenko[2,3], Josef Šivic[1], David Bednář[2,3], Jiří Sedlář[1]

[1]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University in Prague, Jugoslávských partyzánů 1580/3, 160 00 Prague, Czech Republic
[2]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[3]International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno, Czech Republic
jakub.kopko@cvut.cz

Due to the growing recognition of the importance of dynamical properties of proteins [1], molecular dynamics simulations have become an important tool for their analysis. These simulations generate large-scale high-dimensional datasets, motivating the development of data analysis methods aimed at distilling this information into a format understandable to humans. Among these methods, VAMPnet neural network [2] stands out as one of the leading methods, offering a linear perspective on the dynamics by directly learning a Markov state model from the data. The resulting models not only capture information about the metastable states observed during the simulation but also provide the probabilities of transition between these states.

For enhancing the interpretability of Markov state models, we have developed a comparative Markov state analysis approach CoVAMPnet [3]. The CoVAMPnet method introduces a way to assess the significance of inter-residue distances for assigning Markov states through the analysis of aggregated neural network gradients. Additionally, CoVAMPnet enables comparison of Markov states between different sets of simulations, including simulations corresponding to different systems. This is achieved through the alignment of ensembles of Markov state models obtained as a solution to an optimal transport problem. The versatility of CoVAMPnet allows its application in various research areas where the comparison of molecular dynamics simulations of different systems finds its use. Crucially, this includes the study of the effects of drug candidates on the conformational behaviour of intrinsically disordered biomolecules [3] and the analysis and comparison of closely related protein variants, illuminating the influence of the mutations on the dynamical properties of the protein.

We applied CoVAMPnet to analyse the dynamics of apolipoprotein E (APOE), specifically its 4-helix bundle

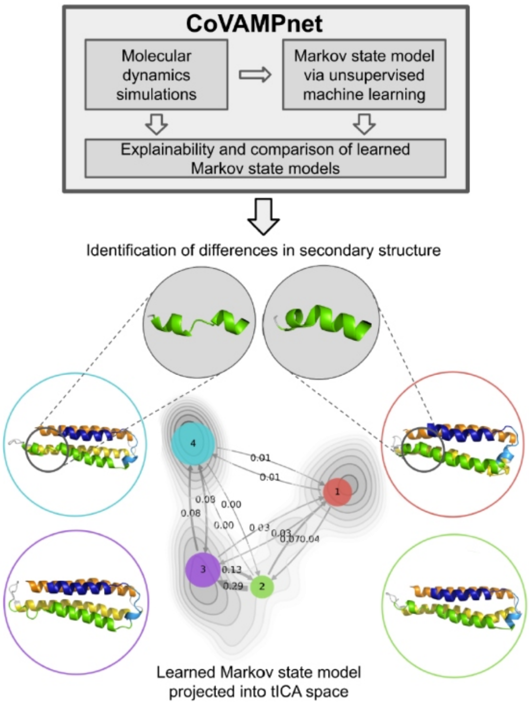

CoVAMPnet
Molecular dynamics simulations
Markov state model via unsupervised machine learning
Explainability and comparison of learned Markov state models
Identification of differences in secondary structure
4
1
3
2
0.01
0.01
0.09
0.00
0.08
0.00
0.03
0.03
0.07
0.04
0.13
0.29
Learned Markov state model projected into tICA space

Figure 1. Overview of the CoVAMPnet method with an example result. The loss of structure in the H2 subdomain (green) of the 4-helix bundle was observed in the simulation of APOE4 with 3SPA. Shortening and extension of helices of APOE have been previously identified experimentally using hydrogen deuterium mass-spectrometry.

domain, which plays a role in the APOE dimerization process [4]. In the first part of our investigation, we focused on differences between APOE3 and APOE4, two APOE variants that appear to exhibit different risk levels of Alzheimer's disease onset in their human carriers. Subsequently, our research also explored the influence of the small molecule drug candidate 3SPA on the conformational behaviour of APOE3 and APOE4, providing valuable insights into possible explanations of its therapeutic potential.

With CoVAMPnet we verified the previously reported findings [4] and gained two interesting novel insights into the flexibility of APOE. Firstly, we found out that the HL1 subdomain of the free APOE3 is remarkably flexible, which we hypothesise may play a role in forming dimer interfaces [4]. Secondly, APOE4 simulated in the presence of 3SPA formed previously unknown conformations with a specific loss of structure in the H2 subdomain. Traditional methods of data processing missed this specific conformational change [4], demonstrating CoVAMPnet's effectiveness in comprehensively analyzing complex protein dynamics.

1. Miller MD, Phillips GN Jr., *Moving beyond static snapshots: Protein dynamics and the Protein Data Bank*. Journal of Biological Chemistry, **296** (2021).

2. Andreas Mardt, Luca Pasquali et al. *VAMPnets for deep learning of molecular kinetics*. Nature Communications, **9**, (2018).

3. Sérgio M. Marques, Petr Kouba et al. Effects of Alzheimer's Disease Drug Candidates on Disordered Aâ42 Dissected by Comparative Markov State Analysis (CoVAMPnet). bioRxiv 2023.01.06.523007, (2023).

4. Michal Nemergut, Sérgio M. Marques et al. *Domino-like effect of C112R mutation on ApoE4 aggregation and its reduction by Alzheimer's disease drug candidate*. Molecular Neurodegeneration, **18**, (2023).

This work was supported by the Technology Agency of the Czech Republic through projects RETEMED and TEREP (TN02000122, TN02000122/001N) and by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254). This research has received funding from The project National Institute for Neurological Research (Programme EXCELES, ID Project No. LX22NPO5107) - Funded by the European Union – Next Generation EU.

P21

DeepREx2.0: PROTEIN LANGUAGE MODELS IMPROVE THE PREDICTION OF RESIDUE SOLVENT ACCESSIBILITY FROM SEQUENCE

M. Manfredi, C. Savojardo, P.L. Martelli, R. Casadio

Biocomputing Group, Dept. of Pharmacy and Biotechnology, University of Bologna, Italy
matteo.manfredi4@unibo.it

Knowledge of residue solvent accessibility in a protein is important for different applications, including the identification of interacting functional surfaces and the characterization of residues undergoing variations. Accessible Surface Area (ASA) and Relative Solvent Accessibility (RSA) values can be directly computed from the 3-dimensional structure. When resolved structures are not available, machine learning-based tools can provide accurate estimates starting from the protein sequence.

Recently introduced Protein Language Models (PLMs) allow the development of faster and more accurate tools with respect to canonical encodings such as Multiple Sequence Alignments (MSAs). After being trained on huge datasets including billions of protein sequences in a self-supervised way, PLMs can be efficiently adopted to generate a representation of the protein sequence that casts relevant evolutionary, structural, and contextual information. Different architectures are then fine-tuned to provide task-specific predictions.

Here we present DeepREx 2.0, a tool to accurately estimate the RSA values for residues of a protein sequence in the absence of the structure.

DeepREx 2.0 is trained on 6,552 proteins and benchmarked on 21 proteins, all obtained from the PDB. Each protein has been mapped on the corresponding UniProt entry and DSSP was adopted to compute accessibility values from protein structures. The training dataset was split into 10 equally sized subsets for performing a 10-fold cross-validation. Proteins included in two different subsets share less than 25% sequence identity over a minimum of 40% coverage. Moreover, the blind test set is completely non-redundant with respect to the training datasets of all methods included in the benchmark.

The method is described in Figure 1 and in the corresponding legend. The input encoding is based on two state-of-the-art PLM-based embeddings, namely ProtT5 [1] and ESM2 [2], and the prediction is performed with a Deep Learning architecture.

The method achieves a 0.72 Pearson Correlation Coefficient on the blind test set, outperforming the first release of DeepREx [3], which exploits evolutionary information contained in MSAs. Moreover, DeepREx 2.0 outperforms two recently released methods based on PLM encoding, NetSurfP-2.0 [4] and SPOT-1D-LM [5]. The peculiarity of DeepREx 2.0 is to adopt two different PLMs to embed the input. Results then confirm our previous observation [6-8] that the concatenation of embeddings generated by different and complementary models can improve the performance of the downstream predictors.

DeepREx 2.0 is available at
https://deeprex.biocomp.unibo.it/v2/

Table 1. Benchmark of DeepREx 2.0 and state-of-the-art methods

Method	PCC	MSE	MCC (T = 20%)
DeepREx 2.0	0.72	0.03	0.63
DeepREx (old) [3]	0.56	0.08	0.49
NetSurfP-3.0 [4]	0.67	0.04	0.58
SPOT-1D-LM [5]	0.71	0.04	0.62

1. Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Y, Jones L, et al. *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing*. arXiv [csLG] 2020.
2. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. *Evolutionary-scale prediction of atomic level protein structure with a language model*. Science. 202; **379**:1123-1130.
3. Manfredi M, Savojardo C, Martelli PL, Casadio R. *DeepREx-WS: A web server for characterising protein–solvent interaction starting from sequence*. Comput Struct Biotechnol J 2021;**19**:5791–9.
4. Hřie MH, Kiehl EN, Petersen B, Nielsen M, Winther O, Nielsen H, et al. *NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning*. Nucleic Acids Res 2022;**50**:W510–5.
5. Singh J, Paliwal K, Litfin T, Singh J, Zhou Y. *Reaching alignment-profile-based accuracy in predicting protein secondary and tertiary structural properties without alignment*. Sci Rep 2022;**12**:7607.
6. Manfredi M, Savojardo C, Martelli PL, Casadio R. *E-SNPs&GO: embedding of protein sequence and function improves the annotation of human pathogenic variants*. Bioinformatics 2022;**38**:5168–74.
7. Manfredi M, Savojardo C, Luigi Martelli P, Casadio R. *ISPRED-SEQ: Deep neural networks and embeddings for predicting interaction sites in protein sequences*. J Mol Biol 2023:167963.
8. Madeo G, Savojardo C, Manfredi M, Martelli PL, Casadio R. *CoCoNat: a novel method based on deep learning for coiled-coil prediction*. Bioinformatics 2023;**39**.

Figure 1. Architecture of DeepREx 2.0. The target sequence is first processed by two different and complementary PLMs, ProtT5 [1] and ESM2 [2], to generate a concatenated vector of 1280+1024=2304 features for each residue. A sliding window of 31 residues is then processed by the network, consisting of a Convolutional layer with 2304 filters followed by a stack of 3 dense networks. The output consists of a single value between 0 and 1, representing the putative RSA of the residue. A threshold of 20% is also adopted to distinguish Buried and Exposed residues.

P22

PROTEOME SECONDARY STRUCTURES GENERATED BY ALPHAFOLD

Ivana Hutařová Vařeková[1], Dominik Martinat[1], Radka Svobodová[2,3], Karel Berka[1]

[1]Department of Physical Chemistry, Faculty of Science, Palacký University, tř. 17. listopadu 12, 77146 Olomouc, Czech Republic
[2]CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
[3]National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno, Czech Republic
dominik.martinat@upol.cz

The AlphaFold [1] algorithm and its associated database [2] provide convenient access to the structural data for entire proteomes across various species, facilitating comprehensive statistical analysis. In our study, we examined the distribution of secondary structures, specifically alpha helices and beta strands, within the proteomes of model organisms and those of relevance to human health. While there seems to be a general distribution of secondary structures within proteins in all analyzed life forms, our findings reveal several noteworthy functional exceptions: 1) abundance of short proteins with small amounts of secondary structures in plant proteomes, 2) spike of structures with 9-12 alpha helix count in some mammals (e.g., mice or rat) from the abundance of olfactory receptors from GPCR family and 3) enhanced presence of long proteins with abundant secondary structures (50+) in the human proteome. These insights contribute to a deeper understanding of the structural diversity within proteomes, shedding light on specific patterns and variations across different species and functional categories of proteins.

1. Jumper J., Evans R., Pritzel A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2.
2. Varadi M., Anyango S., Deshpande M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models, Nucleic Acids Research, Volume 50, Issue D1, D439–D444 (2022), https://doi.org/10.1093/nar/gkab1061.

P23

OLD FOLDS CAN LEARN NEW TRICKS: ALPHAFOLD-DRIVEN INSIGHTS ON COILED-COIL STRUCTURE

M. Martinez-Goikoetxea[1], Rafal Madaj[2], Jan Ludwiczak[2,3], A. Lupas[1], S. Dunin-Horkawicz[1,2]

[1]Department of Protein Evolution, Max Planck Institute for Biology, Tübingen, Germany
[2]Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland
[3]Prescient Design, Genentech Research & Early Development, Roche Group, Basel, Switzerland
mikel.martinez@tuebingen.mpg.de

Coiled coils are a widespread protein structure motif that consists of two or more -helices that wind around a central axis to form a helical bundle with a buried hydrophobic core. They are built from relatively short and simple sequence repeats, typically consisting of 7 residues (heptads), although alternative repeat sizes are possible, the most common being 11 residues (hendecads) [1,2]. The repeat size and residue composition of coiled coils are responsible for their considerable variety in terms of the helical topology (number and orientation of the helices) and geometry (axial rotation and degree and direction of the winding). Conversely, these structural features are responsible for the extraordinary diversity of functions that coiled-coil domains perform in nature, such as mechanical support, muscle contraction, vesicle transport and fusion, transcription factor, or signal transduction. In this work, we have benchmarked how accurate AlphaFold is in modeling typical heptad coiled coils [3], and investigate whether it could be applied to new, hitherto undescribed non-heptad coiled coils such as the ones composed primarily of hendecads. Our results show that AlphaFold is able to recapitulate a number of known coiled-coil rules that relate sequence and structure, and that it can be used to obtain insights into new ones. Simultaneously, we have found a number of cases that highlight the limitations and biases of AlphaFold in coiled-coil modeling. We hope that our work will serve as a foundation to develop new tools with which to further advance our understanding of this model protein structure motif.

1. M. Martinez-Goikoetxea and A. Lupas. "A Conserved Motif Suggests a Common Origin for a Group of Proteins Involved in the Cell Division of Gram-Positive Bacteria." *PloS One* 18, no. 1 (2023): e0273136. https://doi.org/10.1371/journal.pone.0273136.

2. M. Martinez-Goikoetxea and A. Lupas. "New Protein Families with Hendecad Coiled Coils in the Proteome of Life." *Journal of Structural Biology* 215, no. 3 (September 1, 2023): 108007. https://doi.org/10.1016/j.jsb.2023.108007.

3. R. Madaj, M. Martinez-Goikoetxea, J. Ludwiczak, K. Kaminski and S. Dunin-Horkawicz. Manuscript in preparation.

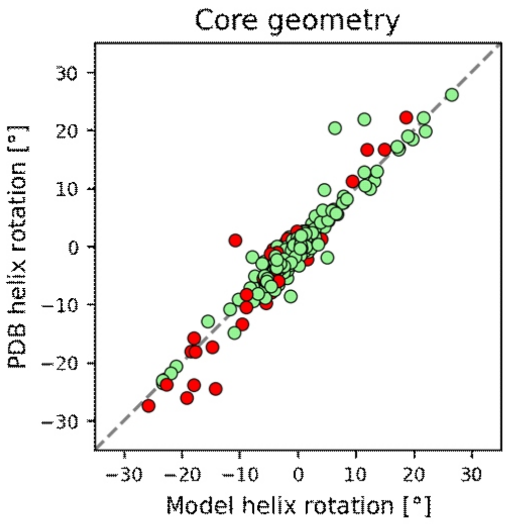


Figure 1. CC core geometry modeling benchmark.

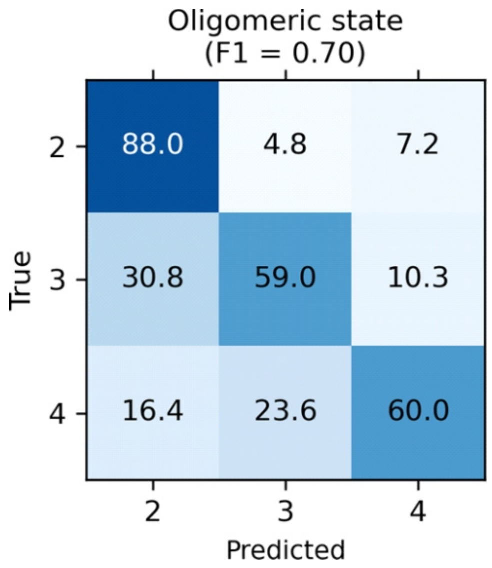


Figure 2. CC oligomeric state benchmark.

P24

THE PROTEIN UNIVERSE ATLAS

J. Durairaj[1,2], A. M. Waterhouse[1,2], T. Mets[3,4], T. Brodiazhenko[3], M. Abdullah[3,4], G. Studer[1,2], G. Tauriello[1,2], M. Akdel[5], A. Andreeva[6], A. Bateman[6], T. Tenson[3], V. Hauryliuk[3,4,7,8], T. Schwede[1,2], J. Pereira[1,2]

[1]Biozentrum, University of Basel, Basel, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Basel, Basel, Switzerland
[3]Institute of Technology, University of Tartu, Tartu, Estonia
[4]Department of Experimental Medical Science, Lund University, Lund, Sweden
[5]VantAI, New York, USA
[6]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
[7]Science for Life Laboratory, Lund, Sweden
[8]Virus Centre, Lund University, Lund, Sweden
joana.pereira@unibas.ch

The collection of all protein sequences sampled by nature is commonly referred to as the “Protein Universe”. This can be seen as a multidimensional space of amino acid sequences, where each protein adopts a coordinate that is defined by its similarity to all others. This landscape can be conceptualized as a large protein similarity network, where protein families and superfamilies form clusters and superclusters whose internal structure illustrates the evolutionary (or other similarity relationships) of the encompassed proteins. While commonly used for the study of individual protein or domain families, the methods currently available are memory intensive and are limited to relatively small sets of up to 10 to 20 thousand proteins.

Leveraging the recent advances in GPU-accelerated force-directed graph layouting and complex network summarizing approaches, we constructed for the first time a protein similarity network encompassing more than 50 million proteins. We annotated this network, linked it to various protein databases and made it available as a new web resource, the Protein Universe Atlas, available at https://uniprot3d.org. In contrast to common protein databases such as UniProt, GenBank and the Protein Data Bank, which are protein-centric and display a page for each single queried protein, the Atlas represents entries as points in a 2D landscape where similar proteins are close in space. This puts them in evolutionary context, highlighting evolutionary relationships that can help guide biocuration, protein family classification, and protein function prediction and annotation.

Thanks to this development, we were already able to shed light onto a new toxin-antitoxin superfamily, to discover a new protein fold family, and to add hundreds of not hitherto described protein families to Pfam. We expect that expanding the resource to the more than 250 million proteins in UniProt and to those resulting from large-scale metagenomic studies will significantly speed up the study of any protein of interest, the automatic identification of protein families and superfamilies, and the functional annotation of the exponentially growing protein data deposited in protein databases.

P25

EXPLORING THE RELATIONSHIP BETWEEN STRUCTURAL QUALITY AND JOURNAL IMPACT FACTOR

Jana Porubská[1], Veronika Horská[1], Vladimír Horský[1], Radka Svobodová[1,2]

[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[2]CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
jana.porubska@mail.muni.cz

In structural bioinformatics, scientific research heavily depends on the structural data stored in the Protein Data Bank (PDB). The scientific community’s widespread utilization of structural models puts pressure on the integrity of these structures. Consequently, a profound emphasis exists on structural validation, with the PDB providing detailed validation reports with various quality attributes.

Our project embarks on an exploration of structural quality across different scientific journals. We investigate the interplay between structural quality metrics and the journal impact factor. Our methodology involves a selection of journals, categorizing them based on their impact factor, and the compilation of a dataset containing selected quality factors of these structures. Finally, our analysis uses statistical methods to describe the dynamic nature of the structural quality within each journal category.

P26

TOWARDS FINDING HIGH AFFINITY BINDERS (PEPTIDES, PEPTIDE MIMETICS) OF THE PSB DOMAIN OF HUMAN COLLAGEN PROLYL 4-HYDROXYLASE, USING FRET AND COMPUTATIONAL APPROACHES COMPLEMENTED BY CALORIMETRIC AND PROTEIN CRYSTALLOGRAPHIC CHARACTERIZATION

M. Mubinur, Rahman[1], Hongmin Tu[2], Bukunmi Adediran[1], Sudarshan Murthy[1], Antti M. Salo[1,2], Outi Lampela[2], Andre Juffer[2], Johanna Myllyharju[1,2], Rik K. Wierenga[1], M. Kristian Koski[1,2]

[1]Faculty of Biochemistry and Molecular Medicine, University of Oulu, Oulu, Finland
[2]Biocenter Oulu, University of Oulu, Oulu, Finland
rik.wierenga@oulu.fi, kristian.koski@oulu.fi

Collagen prolyl 4-hydroxylase (C-P4H) catalyses the hydroxylation of the Y prolines of the XYG-repeat of procollagen. In human there are three isoforms, C-P4H-I, -II and –III. C-P4Hs are tetrameric $_2$ $_2$ enzymes. The -subunit provides the N-terminal dimerization domain, the middle peptide-substrate-binding (PSB) domain and the C-terminal catalytic (CAT) domain. The CAT domain belongs the Fe(II) and 2-oxoglutarate-dependent dioxygenases, which adopt the DSBH-fold and which use molecular oxygen to hydroxylate proline residues. The -subunit is identical to PDI and its function for the catalysis is not yet known. The PSB domain (about 100 residues), unique for the collagen prolyl 4-hydroxylases, binds the proline rich peptide substrate that is hydroxylated by the CAT domain. The PSB domain has five -helices, adopting the TPR fold.

Crystal structures are now available of all domains and subunits of C-P4H. For these structures different truncated forms of the -subunit have been used. It concerns the crystal structure of the dimeric DD construct (of C-P4H-I), consisting of the dimerization and PSB domains of the subunit, showing the dimerization motif between the two subunits [1], whereas the crystal structure of the CAT domain (of the truncated C-P4H-II subunit), complexed with mature PDI (the CAT-PDI complex) shows the mode of interactions of the CAT domain and PDI [2]. These structures do not reveal the mode of assembly of the DD-dimer with the two CAT-PDI units, which is the structure of the mature C-P4H $_2$ $_2$ complex. The latter structure has however been predicted [2] by AI-based structure prediction tools, such as Robetta and AlphaFold.

The PSB domain is important for the efficient hydroxylation of proline rich peptides by the CAT domain [3]. The binding of proline rich peptides to the PSB-I domain has previously been reported for the peptides $(PPG)_3$ and P9, which bind in an extended, PP-II conformation [1]. Protein crystallographic binding studies show that the mode of binding to PSB-I of proline-rich peptides that have the PxGP motif, is different, adopting a bent conformation, which is the same as previously reported for the mode of binding of such peptides to the PSB-II domain [4]. The PSB domain has two proline binding pockets, referred to as the P5 and P8 pockets. In both modes of binding proline side chains bind in these pockets. Calorimetric binding studies show that these PxGP peptides have good affinity for the PSB-I domain.

The current PSB studies are aimed at finding tight binders at the P5 and P8 pockets of the PSB domain. Such compounds will compete with the binding of the proline rich substrate peptides and are therefore potential inhibitors of C-P4Hs. Such compounds are potential pharmaceuticals against fibrotic diseases and cancer. Two approaches are being developed to find these compounds (i) the FRET method using various screens with a wide range of compounds and (ii) biocomputational methods using the currently available structures of PSB peptide complexes as a starting point. The poster will describe the currently available structural and affinity data.

This project is funded by the Jane and Aatos Erkko foundation.

1. Anantharajan, J., Koski, M. K., Kursula, P., Hieta, R, Bergmann, U., Myllyharju, J, Wierenga, R. K., Structure, 21 (2013), 2107-2118.
2. Murthy, A. V., Sulu, R., Lebedev, A., Salo, A. M., Korhonen, K., Venkatesan, R., Tu, H., Bergmann, U., Jänis, J., Laitaoja, M., Ruddock, L., Myllyharju, J., Koski, M. K., Wierenga, R. K., J. Biol. Chem., 298 (2022), 102614.
3. Pekkala, M., Hieta, R., Bergmann, U., Kivirikko, K. I., Wierenga, R.K., Myllyharju, J., J. Biol. Chem., 279 (2004), 52255-52261.
4. Murthy, A. V., Sulu, R., Koski, M. K., Tu, H., Anantharajan, J., Sah-Teli, S. K., Myllyharju, J., Wierenga, R. K., Prot. Sci., 27 (2018),1692-1703.

P27

Onedata4Sci: LIFE-SCIENCE EXPERIMENTAL DATASETS MANAGEMENT SYSTEM

Adrian Rosinec[1,2,3], Tomas Svoboda[1,2,3], Tomas Racek[1,2,3], Josef Handl[1], Jozef Sabo[2,3], Ales Krenek[3], Radka Svobodova[1,2]

[1]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, Brno, 625 00, Czech Republic
[2]CEITEC - Central European Institute of Technology, Kamenice 753/5, Brno, 625 00, Czech Republic
[3]Institute of Computer Science, Masaryk University, Sumavská 416/15, Brno, 602 00, Czech Republic
adrian@muni.cz

In many scientific disciplines, especially life-sciences, expensive equipment is shared nowadays (like cryoEM devices, optical microscopes, …). The users – scientists request specific experiments from facilities, which perform the experiments on their behalf. The outcome of such an experiment is a dataset, which can get quite large in many cases (tens of gigabytes to terabytes). Data are being processed in order to draw scientific conclusions by their interpretation, then results are published. Nowadays, emphasis is given to the availability of the data so that any scientific results can be verified independently making datasets FAIR. Managing the whole life-cycle of those data can be tedious, especially when data are being acquired in a period of time.

To address these challenges, we design and develop a system Onedat4Sci, that automates acquiring, sharing, and publishing of data produced by specialized scientific devices. The proposed solution automatically makes experimental data available to the scientific community in a predefined way. It is particularly useful for on-the-fly processing in local or distant data centers, real-time analysis, or archiving to permanent storage according to defined quality of service (e.g., data distribution). The solution includes a web-based system that can be used to manage emerging datasets and annotate them with metadata (automatically extracted from the data produced by the instruments or manually entered by users according to defined templates).

P28

THE ANALYSIS OF THE IMPACT OF SUBSTITUTIONS WITHIN EXO-MOTIFS ON Hsa-MiR-1246 INTERCELLULAR TRANSFER IN BREAST CANCER CELLS

A. Rybarczyk[1,2], T. Lehmann[3], E. Iwańczyk-Skalska [3], W. Juzwa[4], K. Kopciuch[1], P.P. Jagodziński[3]

[1]Institute of Computing Science, Poznan University of Technology, Poland
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland
[3]Department of Biochemistry and Molecular Biology, Poznan University of Medical Sciences, Poznan, Poland
[4]Department of Biotechnology and Food Microbiology, Poznan University of Life Sciences, Poznan, Poland
arybarczyk@cs.put.poznan.pl

Extracellular vesicles (EVs) release various biomolecules into the extracellular space, with miR-1246 garnering recent interest for its oncogenic role in several cancers. The processes and mechanisms guiding miR-1246 into EVs and its stability remain elusive. In our study, we explored the influence of single-nucleotide alterations in miR-1246's exosome sorting motifs (EXO-motifs: GGAG and GCAG) through both in silico methods, such as structural analysis and modeling, and in vitro assays involving the transfection of fluorescently labeled miRNA to MDA-MB-231 cells, which we analyzed by flow cytometry and fluorescent microscopy. Our findings indicate that disruptions in miR-1246 EXO-motifs can affect its stability and intercellular transfer, suggesting a connection between RNA stability and EV-mediated transfer.

P29

THE HIJACKER OF HOST IMMUNE SYSTEM: STRUCTURAL ANALYSIS OF *KLEBSIELLA* IMMUNE EVASIN A

Mia Åstrand[1], Tobias Eriksberg[1], Joana Sá-Pessoa[2], José A Bengoechea[2], Christine Touma[1],Tiina A. Salminen[1]

[1]Structural Bioinformatics Laboratory and InFLAMES Research Flagship Center, Biochemistry, Faculty of Science and Engineering, A°bo Akademi University, Turku, Finland
[2]Wellcome-Wolfson Institute for Experimental Medicine, School of Medicine, Dentistry and Biomedical Sciences, Queen's University Belfast, Belfast, UK
Tiina.Salminen@abo.fi

The multidrug resistant bacteria *Klebsiella pneumoniae,* which causes infections in the respiratory system and the urinary tract, as well as life-threatening hospital-acquired infections, is included on WHO's list of pathogens that need to be prioritized when developing new antibiotics. Due to the increasing spread of antibiotic resistance among its strains it is considered an urgent problem. The recently discovered *K. pneumoniae* protein, *Klebsiella* immune evasin A (KivA), has been shown to inhibit the IL-17 and TLR signaling pathways that play a significant role in the human immune response against pathogens. KivA consists of an N-terminal SEFIR domain, which is similar to the eukaryotic SEFIR domain, and a C-terminal domain with an unknown 3D fold. KivA inhibits key proteins in the signaling pathways by forming SEFIR-SEFIR interactions with them, but nothing is known about the role of the C-terminal domain yet. We have used traditional and AI-based methods to model the 3D structure for KivA and crystal structure determination to validate the modeling results is ongoing. The SEFIR domain of KivA is predicted to have a typical a flavodoxin-like fold whereas the C-terminal domain forms a novel alpha-helical 3D fold. The structural and functional analysis of KivA will provide insight into the mechanisms used by *K. pneumoniae* to disrupt the immune responses and will enable the development of new effective and more selective treatment methods.

P30

ChannelsDB 2.0: A COMPREHENSIVE DATABASE OF PROTEIN TUNNELS AND PORES IN ALPHAFOLD ERA

Anna Špačková[1], Ondřej Vávra[2,3], Tomáš Raček[4,5], Václav Bazgier[1], David Sehnal[4,5], Jiří Damborský[2,3], Radka Svobodová[4,5], David Bednář[2,3], Karel Berka[1]

[1]Department of Physical Chemistry, Faculty of Science, Palacký University, tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
[2]Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[3]International Clinical Research Center, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno, Czech Republic
[4]CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
[5]National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno, Czech Republic; anna.spackova@upol.cz

ChannelsDB 2.0 has been upgraded to offer a comprehensive insight into protein channels' structural characteristics, geometry, and physicochemical attributes, encompassing both tunnels and pores. These channels are computed on deposited biomacromolecular structures originating from the PDBe and AlphaFoldDB databases. In the new version of the ChannelsDB database, we have incorporated data generated through the widely used CAVER tool, augmenting the insights previously acquired only through the original MOLE tool. Additionally, we have extended the database's coverage by introducing tunnels originating from cofactors localised within the AlphaFill database or from cognate ligands within PDB structures. This expansion has increased the available channel annotations by almost five times. ChannelsDB 2.0 houses information concerning geometric properties such as length and radius and physicochemical attributes based on the amino acids lining the channels. These stored data are intricately linked with the existing UniProt mutation annotation data, facilitating in-depth investigations into the functional roles of biomacromolecular tunnels and pores.

In summary, ChannelsDB 2.0 represents an invaluable resource for conducting in-depth analyses of the significance of biomacromolecular channels. The database is freely accessible to the public at the address https://channelsdb2.biodata.ceitec.cz.

P31

C-RCPred: A MULTI-OBJECTIVE ALGORITHM FOR INTERACTIVE SECONDARY STRUCTURE PREDICTION OF RNA COMPLEXES INTEGRATING USER KNOWLEDGE AND SHAPE DATA

Mandy Ibéné, Audrey Legendre, Guillaume Postic, Eric Angel, Fariza Tahi

Université Paris-Saclay, Univ. Evry, IBISC, 91200, Evry-Courcouronnes, France
fariza.tahi@univ-evry.fr

RNAs can bind and form complexes that have important catalytic functions. One can cite, for example, the ribosome, composed of 5S, 5.8S, 18S and 28S RNAs in eukaryotes, whose role is the translation of messenger RNAs into proteins. The ribosome is also made up of proteins, but it is the RNAs that are responsible for the catalytic activity of the complex. There are also RNA-only complexes, such as the cyclic hexamer of bacteriophage , used to infect bacteria [1].

RNA complexes can adopt secondary and tertiary structures, expressing their biological function. The prediction of their structure is therefore an important issue, as for RNAs alone. In this study, we address the secondary structure prediction issue. Currently, many bioinformatics tools are available for RNA secondary structure prediction (prediction of the structure of an RNA as in [2-4]), as well as for RNA-RNA interaction (interaction between two RNAs as in [4-6], sometimes including the global structure [5]). But very few tools exist for predicting the structure of complexes composed of more than two RNAs. We cite MultiRNAFold [7], NUPACK [8], RCPred [9], RNAmultifold [10] and VfoldMCPX [11].

Biologists may have knowledge that can help the prediction, like sporadic information on the structure: pairings, particular motifs, etc. They may also have their own experimental data like SHAPE data [12-14], which give probing information on the considered RNAs. The integration of user knowledge and probing data into the prediction of secondary structures has been used for RNAs alone [13, 15], as well as for RNA-RNA interaction [16]. However, to the best of our knowledge, it is not the case for RNA complexes.

We have developed a new method called C-RCPred [17] for predicting (in an interactive way) secondary structures of RNA complexes, which allows to integrate probing data and user knowledge. C-RCPred is based on a multi-objective approach where the objectives are the free energy, user constraints and probing data. It takes as input a set of secondary structures per RNA sequence and a set of RNA-RNA interactions per pair of RNAs, which are predicted by existing tools (or provided by the user), and aims to find the best combinations of these inputs, i.e. the combinations leading to complexes optimizing simultaneously the different objectives. For this purpose, it solves a multi-objective weighted graph optimization problem, the *clique problem*. Experimental probing data for some RNAs can be found in databases. However, to our knowledge, there is no specific probing data for RNAs involved in complexes and very few data are available. A recent tool, called Shaker [18], allows to predict SHAPE data for RNAs. We therefore integrated this tool in our multi-objective method.

The performed benchmarks show the efficiency of the multi-objective approach, and the positive impact of considering user knowledge and probing data on the prediction results. They also show C-RCPred gives better prediction results compared with the state-of-the-art methods. The figure below (Fig. 1) shows the predictions obtained by our tool and the state-of-the-art tools on an example of RNA complex.

C-RCPred is an interactive tool, freely available on our bioinformatics platform

EvryRNA: http://evryrna.ibisc.univ-evry.fr

1. Feng Zhang, Sébastien Lemieux, Xiling Wu et al. Function of hexameric RNA in packaging of bacteriophage 29 DNA in vitro. Molecular cell, 2(1):141–147, 1998.
2. Stefan Janssen and Robert Giegerich. The RNA shapes studio. Bioinformatics, 31(3):423–425, 2014.
3. Audrey Legendre, Eric Angel, and Fariza Tahi. Bi-objective integer programming for RNA secondary structure prediction with pseudoknots. BMC bioinformatics, 19(1):13, 2018.
4. Ronny Lorenz, Stephan H Bernhart, Christian Hoener Zu Siederdissen et al. Viennarna package 2.0. Algorithms for Molecular Biology, 6(1):1, 2011.
5. Anke Busch, Andreas S Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics, 24(24):2849–2856, 2008.
6. Hakim Tafer and Ivo L Hofacker. RNAplex: a fast tool for RNA– RNA interaction search. Bioinformatics, 24(22):2657–2663, 2008.
7. Saad Mneimneh and Syed Ali Ahmed. Gibbs/MCMC sampling for multiple RNA interaction with sub-optimal solutions. In International Conference on Algorithms for Computational Biology, pages 78–90. Springer, 2016.
8. Weitian Tong, Randy Goebel, Tian Liu and al. Approximating the maximum multiple RNA interaction problem. Theoretical Computer Science, 556:63–70, 2014.
9. Audrey Legendre, Eric Angel, and Fariza Tahi. RCPred: RNA complex prediction as a constrained maximum weight clique problem. BMC bioinformatics, 20(3):128, 2019.
10. Ronny Lorenz, Christoph Flamm, Ivo L. Hofacker et al. Efficient computation of base-pairing probabilities in

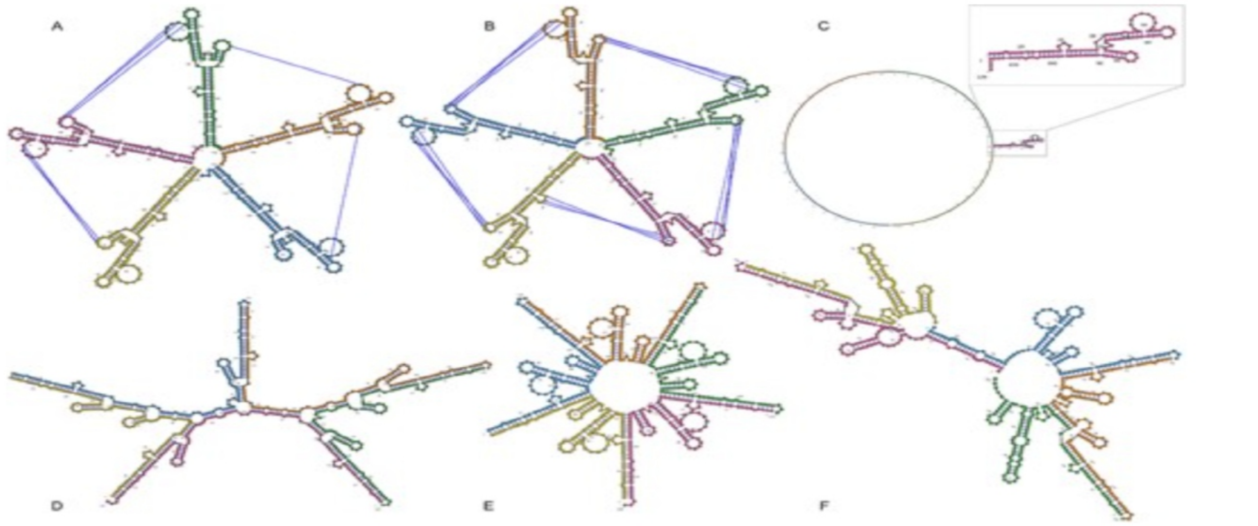


Figure 1. Secondary structure of the pentameric prohead RNA of the bacteriophage 29 (PDB code 1FOQ): comparison between the reference structure (A) and predictions generated by C-RCPred (B), RCPred (C), MultiRNAFold (D), RNAmultifold (E), NUPACK (F). The structures shown are the solutions with the best MCC.

multi- strand RNA folding. In Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, pages 23–31. SCITEPRESS, 2020.

11. S Zhang, Y Cheng, P Guo et al. VfoldMCPX: predicting multistrand RNA complexes. RNA, 28(4):596–608, 2022.
12. Edwards J Merino, Kevin A Wilkinson, Jennifer L Coughlan et al. RNA structure analysis at single nucleotide resolution by selective 2 -hydroxyl acylation and primer extension (SHAPE). Journal of the American Chemical Society, 127(12):4223–4231, 2005.
13. Katherine E Deigan, Tian W LI, David H Mathews et al. Accurate SHAPE-directed RNA structure determination. Proceedings of the National Academy of Sciences, 106(1):97– 102, 2009.
14. Eckart Bindewald, Michaela Wendeler, Michal Legiewicz et al. Correlating SHAPE signatures with three-dimensional RNA structures. RNA, 2011.
15. Ronny Lorenz, Dominik Luntzer, Ivo L. Hofacker et al.SHAPE directed RNA folding. Bioinform., 32(1):145–147, 2016.
16. Milad Miladi, Soheila Montaseri, Rolf Backofen et al.Integration of accessibility data from structure probing into RNA-RNA interaction prediction. Bioinform., 35(16):2862–2864, 2019.
17. Mandy Ibéné, Audrey Legendre, Guillaume Postic, Eric Angel, Fariza Tahi. Brief. in Bioinformatics, 24(4), 2023. https://doi.org/10.1093/bib/bbad225
18. Stefan Mautner, Soheila Montaseri, Milad Miladi et al.ShaKer: RNA SHAPE prediction using graph kernel. Bioinformatics, 35(14):i354–i359, 07 2019.

P32

REUSING AND ACCESSING COMPUTED STRUCTURE MODELS WITH SWISS-MODEL AND MODELARCHIVE

Gerardo Tauriello[1,2], Gabriel Studer[1,2], Stefan Bienert[1,2], Andrew Waterhouse[1,2], Torsten Schwede[1,2]

[1]Biozentrum, University of Basel, Basel, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Basel, Basel, Switzerland
gerardo.tauriello@unibas.ch

A broad spectrum of applications in life science research benefits from the availability of three-dimensional structures of proteins as they provide valuable molecular insights into their function. While experimental structures are still the most accurate option when available, it has now become possible to generate high quality computed structure models for almost every known protein. We are hence in a situation where creating a structural model is not the challenge any more but we can instead aim to make best use of available models and to get the critical details right for a given downstream application.

Within SWISS-MODEL we have therefore expanded our modelling capabilities to take advantage of the available high quality models in the AlphaFold database (AFDB). We now query AFDB models, rank them together with available experimental structures, and use them as templates for modelling. Since the AFDB contains models for more than 200 million proteins and our users create approximately two new SWISS-MODEL projects every minute, we implemented efficient storage of the AFDB models using the OpenStructure Minimal Format (OMF) and a fast and resource-efficient sequence search based on k-mer search which can query the AFDB in seconds.

While the AFDB contains models for large numbers of proteins, it is limited to sequences in UniProt, to monomer models and to the use of the default AlphaFold2 modelling pipeline. Different modelling pipelines have distinct advantages and disadvantages and users can thus benefit from access to a variety of models to find the one which is best suited for their application. Models for any protein and from any modelling pipeline which is not based on experimental data can be stored in the ModelArchive. There, we have seen a large increase of depositions in the past years from 2’770 publicly released models at the end of 2021 to over 74’000 models in September 2023. This increase in depositions was enabled by the establishment of the ModelCIF data format to store models and their metadata and complemented by the development of the 3D-Beacons network which enables sharing of models across multiple model providers.

Taken together, our latest developments in ModelArchive make a variety of computed structure models accessible and SWISS-MODEL can reuse models to help users obtain the best ones for their applications.

P33

PROTEIN FOLDING ENHANCED BY ADVERSARIAL AUTOENCODERS

Guglielmo Tedeschi[1], Vladimír Višňovský[2], Aleš Křenek[2], Vojtech Spiwok[1]

[1]Department of Biochemistry and Microbiology - University of Chemistry and Technology, Prague
[2]Department of Machine Learning and Data Processing - Faculty of Informatics - Masaryk University, Brno
tedeschg@vscht.cz

This research is motivated by acceleration. Molecular simulations make true the possibility to simulate the motion from small molecules to big proteins and their combinations in drug-target complexes. It lets us predict their changing confirmation, their stability and a plenty of other properties thanks to the evolution of molecular structure. However, application of molecular simulations is affected by the large computational costs in computing steps that must be in order of femtoseconds, to assure numerical stability to integrate Newton equation of motion. Taking into account this limitation, a typical molecular dynamics simulation is capable of sampling only a small fraction of the states available to the simulated system, with the likely catch or unlikely loss of some slow or rarely occurring processes, where likelihood depends on the simulation time. There are numerous techniques to address this limitation and to speed up simulations. Metadynamics is an enhancing method based on biasing Hamiltonian of the system that helps to cross barriers and go head through new unexplored free energy surface areas, thanks to some selected internal coordinates, so called collective variables. Choosing correct collective variables to make metadynamics successful is not a trivial task and it depends first of all on the knowledge and expertise of the user. In the last few years there are emerging opportunities for machine learning and artificial neural networks in this field. We decided to develop an adversarial autoencoder [1] as a tool to analyze simulation data and to support users to derive good collective variables to enhance molecular dynamics simulation. The potential of this platform is demonstrated by using Trp-cage and Villin headpiece unbiased molecular dynamics simulations [2]. We aim to explore its applicability in more complex systems.

We thank D.E Shaw Research to provide us trpcage trajectory for asmsa training. This work was supported by the Czech Science Foundation (22-29667S).

1. A. Makhzani, J.Shlens, N.Jaitly and I.Goodfellow, I. Adversarial autoencoders. International Conference on Learning Representations, 2016.
2. K. Lindorff-Larsen, S. Piana, R. Dror and D. Shaw, How fast-folding proteins fold. Science 334:517, 2011.

P34

LIGYSIS: A PIPELINE AND WEB APPLICATION FOR THE ANALYSIS OF LIGAND BINDING SITES

Javier S. Utgés, Stuart A. MacGowan, Geoffrey J. Barton

University of Dundee, United Kingdom
2394007@dundee.ac.uk

Ligands are key for protein function and can act as substrates, co-factors, or inhibitors in complex with a myriad of proteins spread across all molecular processes. For that reason, understanding how they interact with proteins is crucial and can provide insight into drug development, and wider protein function understanding. In this work, we gathered >25,000 proteins from multiple species with experimentally resolved 3D structures. We analyse the interactions between these proteins and biologically meaningful ligands as defined by BioLiP. Protein-ligand interaction fingerprints are used to group ligands and define »100,000 unique ligand binding sites. Ligand sites are grouped in four clusters based on their solvent accessibility profile. The defined clusters are biologically different, in terms of evolutionary divergence, enrichment in neutral missense variation, relative solvent accessibility, and functional enrichment. These results strongly suggest that these cluster labels can be used to infer ligand binding site functionality. These findings will be of interest to those studying protein-ligand interactions or developing new drugs. To facilitate access to these data, we are developing web application that will allow users to explore in structure the defined binding sites on a protein of interest, as well as dynamically explore graphs and tables displaying the features of the residues forming the sites. The LIGYSIS web service is a Python Flask application that uses 3Dmol.js for protein visualisation, Chart.js for dynamic graph rendering, Bootstrap for stylings, and vanilla JavaScript and jQuery to link all the components together and make them interact.

P35

DEFINING CONFORMATIONAL STATES AND THEIR VARIABILITY

Jose Gavalda Garcia, David Bickel, Joel Roca Martinez, Wim Vranken

Interuniversity Institut of Bioinformatics in Brussels, Vrije Universiteit Brussel, Belgium
wim.vranken@vub.be

The dynamics and related conformational changes of proteins are often essential for their function, but are difficult to characterise and interpret. The energy landscape that determines the conformational behavior and dynamics of an amino acid residue in a protein is determined by its local environment, which encompasses interactions with other residues or molecules as well as parameters such as temperature or pH. The lowest energy state for a given residue can correspond to very sharply defined conformations, for example when the residue is part of a stable helix, or can cover a wide range of conformations, for example for residues in intrinsically disordered regions. Defining such low energy states will therefore help to describe the behavior of a residue and how it changes with its environment. We propose a novel data-driven probabilistic definition of six low energy conformational states typically accessible for amino acid residues in proteins. This definition is based on in solution NMR information for 1414 proteins, through a combined analysis of structure ensembles with interpreted chemical shifts. We further introduce a conformational state variability parameter that captures, based on an ensemble of protein structures from molecular dynamics or other methods, how often a residue moves between these conformational states. The approach enables a different perspective on the conformational behavior of proteins that is complementary to their static interpretation from single structure models.

P36

COMPUTATIONAL DESIGN OF A CYCLIC PEPTIDE FOR INHIBITION OF CD59 USING HOTSPOT EXTENSION IN ROSETTA

Max Beining[1,2], Abdulrahim Altoam[1], Jens Meiler[1,3], Christina Lamers[1], Clara T. Schoeder[1]

[1]Institute for Drug Discovery, Leipzig University Faculty of Medicine, Leipzig, Germany
[2]School of Embedded Composite Artificial Intelligence (SECAI), Cooperation of University Leipzig and TU Dresden
[3]Department of Chemistry, Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, United States
beining@izbi.uni-leipzig.de

CD59 is a crucial membrane complement regulatory protein that plays a pivotal role in limiting the activation of the complement system and preventing the formation of the membrane attack complex (MAC) on cell membranes. CD59 restricts the construction of the MAC on cell membranes by inhibiting the transmembrane channel-forming activity of homologous C8 and C9 proteins during the final stage of MAC formation. On various tumor cells, CD59 showed an increased expression and is associated with reduced survival in cancer patients. It has been shown that inhibition of CD59 expression on leukemia cancer cells leads to increased therapy efficacy with the monoclonal antibody Rituximab [1]. In this *early-stage* project, a computational approach was used to address the challenge of designing a potential inhibitor for CD59. By extending critical interaction anchor residues in the C9-CD59 interface, we build de-novo cyclic peptides containing non-canonical amino acids (NCAA) using the generalized kinematic loop closure method in the software suite Rosetta. Docking experiments showed funneling for some designed peptides into the binding region. Also, the creation of bicyclic peptides using an NCAA Cys-Maleimide linker was performed as a second computational design experiment. Promising candidates will be tested in the laboratory to validate their specific binding affinity.

1. Geller, A., & Yan, J. (2019). The Role of Membrane Bound Complement Regulatory Proteins in Tumor Development and Cancer Immunotherapy. In Frontiers in Immunology (Vol. 10). Frontiers Media SA. https://doi.org/10.3389/fimmu.2019.01074

Computations for this work were done using resources of Leipzig University Computing Center. BMBF (Federal Ministry of Education and Research) partly supports this work in DAAD project 57616814 (SECAI, School of Embedded Composite AI, https://secai.org/).

P37

VALIDATION OF NUCLEIC ACID VALENCE GEOMETRY AT DNATCO.DATMOS.ORG

Jiří Černý, Paulína Božíková, Barbora Schramlová, Bohdan Schneider, Lada Biedermannová, Michal Malý

Institute of Biotechnology of the Czech Academy of Sciences, Czech Republic; barbora.schramlova@ibt.cas.cz

Nucleic acids play a fundamental role in all living organisms. They store genetic information and control the process of protein synthesis. For their study, it is essential to determine their spatial arrangement. A better understanding of the structure will help create more accurate structural models that can help explain many biological mechanisms and develop new substances. Several methods are used to construct a threedimensional atomic model of nucleic acid structure. The most common of these is X-ray crystallography. The initial model needs to be further refined, usually performing a number of refinement and validation cycles.

The goal of our work is to characterize the covalent geometry parameters of nucleic acids and to propose new procedures for their validation. We will present our recently proposed approach based on analysis of a sequentially non-redundant, high resolution, quality-filtered reference set of nucleic acid structures. It will provide an intuitive valence geometry validation score to the authors of nucleic acid structures. The reference implementation is currently available at the https://dnatco.datmos.org web address.