SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

Studying the clinical significance of oncopathological gene variants and fusions using *in silico* protein modeling

by Kristóf Madarász

Supervisor: Attila Mokánszki, Ph.D.

UNIVERSITY OF DEBRECEN
DOCTORAL SCHOOL OF CLINICAL MEDICINE

DEBRECEN, 2025

Studying the clinical significance of oncopathological gene variants and fusions using *in silico* protein modeling

by Kristóf Madarász, MSc degree

Supervisor: Attila Mokánszki, Ph.D.

Doctoral School of Clinical Medicine, University of Debrecen

Head of the **Defense Committee**: Németh Norbert, D.Sc.

Reviewers: Csősz Éva, D.Sc.

Vásárhelyi Barna, D.Sc.

Members of the Defense Committee:

Balogh István, D.Sc.

Czajlik András, Ph.D

The PhD Defense takes place at the Lecture Hall of Bldg. A, Department of Internal Medicine,

Faculty of Medicine, University of Debrecen, 7th October 2025, 1. p.m

1. Introduction

Myelodysplastic syndromes (MDS) and acute myeloid leukemia (AML) are hematological malignancies arising from clonal disorders of the hematopoietic system, characterized by genomic instability and impaired differentiation of myeloid stem cells. MDS comprises a heterogeneous group of diseases defined by ineffective hematopoiesis and peripheral blood cytopenias, whereas AML is a rapidly progressing, aggressive disease dominated by the uncontrolled proliferation of leukemic blasts. The close relationship between the two conditions is confirmed by the fact that a significant portion of MDS cases progress to AML.

Globally, the incidence of both diseases increases with age, particularly in the population over 70 years old. From a genetic perspective, mutations in genes regulating DNA methylation, chromatin remodeling, and RNA splicing play a prominent role in pathogenesis, promoting clonal evolution and therapeutic resistance. The prognosis for subgroups characterized by a complex karyotype and TP53 mutations is particularly unfavorable, with a short median survival time even with intensive treatment, especially in cases of biallelic mutations or high variant allele frequency.

Modern prognostic models, such as the Molecular International Prognostic Scoring System (IPSS-M), integrate clinical parameters, cytogenetic abnormalities, and somatic mutations, thereby refining patient risk stratification. This approach emphasizes the decisive role of the genetic profile in treatment decisions.

One of the greatest challenges in oncological research is the interpretation of variants of uncertain significance (VUS). With the spread of next-generation sequencing (NGS) technologies, an increasing number of genetic alterations are being identified, whose functional and clinical effects are often unclear. The precise evaluation of these variants is a fundamental prerequisite for the application of effective targeted therapies and the success of precision oncology, which requires an integrated approach, including the use of bioinformatics prediction tools.

Gene fusions also play a prominent role in the pathogenesis of many tumors, including sarcomas. These alterations create oncogenic proteins that can be effective targets for targeted therapies. The development of NGS and bioinformatics tools has enabled more accurate detection of these fusions, but interpreting their clinical significance and functional consequences remains a challenge. The integrated analysis of multi-omics data is essential for mapping the signaling pathways activated by these fusions and for identifying new therapeutic targets.

2. Aims

Our work aims to study two fundamental molecular genetic alterations underlying malignant transformations: smaller-scale aberrations affecting nucleotide bases (SNVs, indels) and larger-scale gene fusions (translocations) that result in structural changes. To achieve this, we plan to investigate TP53 gene variants, one of the most common alterations leading to oncohematological diseases, and BCOR gene rearrangements, which occur in some of the rarest soft tissue sarcomas. We intend to analyze the protein-level effects of nucleotide alterations and gene fusions of unknown clinical significance, detected through next-generation sequencing, using *in silico* bioinformatics programs.

I. Investigation of the clinical and molecular significance of *TP53* gene mutations in patients with myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML).

- To investigate the clonal heterogeneity of *TP53* gene mutations in MDS and AML bone marrow samples.
- To examine the correlations between *TP53* mutational status and the severity of hematopoietic disorders.
- *In silico* analysis of the structural and functional changes in the TP53 protein caused by *TP53* gene mutations, with a special focus on protein-protein and protein-DNA interactions.
- Comparative evaluation of different pathogenicity scoring systems to facilitate the interpretation of *TP53* gene mutations.
- To perform *in silico* sequence- and structure-based analyses to investigate changes in TP53 protein stability.

II. Elucidation of the oncogenic mechanism of BCOR-rearranged sarcomas (BRS) using *in silico* approaches.

- To collect available literature data on the BRS patient group, including cDNA sequences of fusion genes based on reverse transcription quantitative PCR (RT-qPCR) and sequencing data, and to compile the amino acid sequences of the fusion proteins.
- Applying *silico* approaches to analyze the sequences and structures of fusion proteins, focusing on changes in functional domains and protein-protein interactions.
- To construct 3D structures of fusion proteins to study the oncogenic mechanism of sarcomas with BCOR fusions, including changes in the binding affinity of the RAWUL-PUFD domains and their effects on interactions within the PRC1 complex.

1. Methods and materials

1.1 Molecular and *in silico* protein analysis of *TP53* mutations in myelodysplastic neoplasia and acute myeloid leukemia

1.1.1 Patients and Samples

Patients were treated at the Department of Hematology, Clinical Center, University of Debrecen. Formalin-fixed, paraffin-embedded (FFPE) bone marrow biopsy tissue samples were retrospectively analyzed from a total of 77 patients who had been reclassified according to the WHO 2022 guidelines into AML-MR (acute myeloid leukemia with myelodysplasia-related changes), myelodysplastic syndrome with increased blasts (MDS-IB; 12 cases), or myelodysplastic syndrome with low blasts (MDS-LB; 39 cases) at the Institute of Pathology, Clinical Center, University of Debrecen. Hematoxylin and eosin (H&E) stained slides were categorized by a pathologist. Cytogenetic analysis was performed as part of the routine diagnostic procedure. We had ethical approvals for the studies (60355-2/2016/EKU and IV/8465-3/2021/EKU). The studies were conducted by the guidelines of the Declaration of Helsinki.

1.1.2 Immunohistochemical studies

Following the examination of H&E slides, TP53 IHC analysis was performed using the Do-07 clone (Dako, Agilent Technologies Company, Santa Clara, CA, USA). IHC positivity was determined if the TP53 staining intensity was high (3+) and at least 10% of the cells were positive.

1.1.3 DNA isolation from FFPE samples

For the extraction of genomic DNA (gDNA) from FFPE tissues, the QIAamp DNA FFPE Tissue Kit (Qiagen, Hilden, Germany) was used. The isolation was performed according to the manufacturer's instructions, and the gDNA was dissolved in 50 µL of elution buffer. DNA concentrations were measured using the Qubit dsDNA HS Assay Kit with a Qubit 4.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

1.1.4 Next-generation sequencing

Following gDNA fragmentation, libraries were prepared using the Accel-Amplicon Comprehensive TP53 panel (Swift Biosciences, Ann Arbor, MI, USA). Sequencing was performed on a MiSeq system (MiSeq Reagent Kit v3 600 cycles, Illumina, San Diego, CA, USA). The libraries (final concentration: 4 nM, pooled at equal molarity) were denatured with 0.2 nM NaOH and then diluted to 40 pM with Illumina hybridization buffer (San Diego, CA, USA). The final loading concentration was 10 pM library and 5% PhiX. Sequencing was performed according to the MiSeq user manual. The prepared libraries were sequenced in a multiplex format with a paired-end run to obtain 2 × 150 bp reads with at least 250X coverage. Adapter-trimmed fastq files were generated using MiSeq Reporter (Illumina, San Diego, CA, USA). Raw sequence data were analyzed for the presence of SNVs and indels using NextGENe software (v.2.4.3; SoftGenetics, State College, PA, USA). The GRCh37 (UCSC hg19) reference genome was used for sequence alignment. The cutoff for individual variants was 5% VAF.

1.1.5 *In silico* protein analysis

Information regarding the TP53 protein (P04637, P53_HUMAN) was collected from the UniProt database and the RCSB Protein Data Bank. The 3D structures of truncated proteins were constructed using the Robetta software. Post-translational modification (PTM) sites of the TP53 protein affected by mutations were identified using the PhosphoSite Plus website. The structural disorder of sequences was determined using the IUPred3 web server. Secondary structure was estimated using the GORIV application. Sequence-based (Seq) stability changes of variants were determined using the I-Mutant2.0 and DDGun servers with default parameters. For structure-based (struc) stability tests, the I-Mutant2.0, DynaMut2, and DDGun3D applications were used. Predictions were made using several TP53 crystal structures (Cry) (5MCT, 5MG7, 5MF7, 1AIE, 1C26, 2FOO, 1YC5) and 3D models built by the AlphaFold (AF) high-accuracy deep learning algorithm. The predicted effect of mutations on protein stability ($\Delta\Delta G_{stability}$ value) was compared with experimentally determined differences for the p.G245S and p.R248Q mutations, as experimental data were only available for these cases.

For the analysis of protein-protein interactions (PPI), the mCSM-PPI2 program was used to calculate the changes in the interaction of TP53 monomers in a homotetramer structure for the p.G334R mutation, and to determine the interaction changes involving the p.S362N mutation during interaction with the ubiquitin-carboxyl-terminal hydrolase 7 (or herpesvirus-associated ubiquitin-specific protease; USP7/HAUSP) protein. For the PPI analysis of the p.G334R mutation, the 2FOO crystal structure was used as it contained the N-terminal domain of USP7/HAUSP in complex with a TP53 protein. For the p.S362N

mutation, crystal structures containing the tetrameric oligomerization domain of TP53 were used: 1OLG and 1SAL. Changes in the affinity of mutant TP53 proteins for DNA were calculated using mCSM-NA.

To determine the pathogenicity of variant proteins, the National Cancer Institute International Agency for Research on Cancer (IARC) TP53 database (R20, July 2019) and its components (Align-GVGD, BayesDel, REVEL, Sift class, PolyPhen2, Transactivation, TransactivationClass, DNE_LOFclass, DNE class, Structure/Function class), as well as the ClinVar, Varity, Phd-SNPg, and FATHMM-XF pathogenicity scoring systems, were used.

1.2 *In silico* analysis of *BCOR*-rearrangement sarcomas

1.2.1 Protein information

We investigated the nucleotide sequences of nine fusions involving the BCOR gene known in the literature. The Ensembl database (GRCh37.p13; GCA_000001405.14) was used to locate the fusion breakpoints and manually assemble the sequences. The coding amino acid sequences were determined in silico using ExPASy: Translate. Information on the BCOR/BCOR-2, CCNB3, MAML3, ZC3H7B, KMT2D, CIITA, RTL9, CLGN, MAML1, and AHR proteins, as well as the assembled fusion proteins, was obtained from the UniProt database (gene and protein fusions were denoted by a double colon (::) according to the latest nomenclature). Domain information was retrieved from the UniProt, Conserved Domain Database (CDD), and InterPro (version 5.67-99.0) databases. Gene Ontology (GO) terms were determined based on the assembled amino acid sequences using InterPro, specifically applying PANTHER GO terms.

1.2.2 Calculation of physicochemical properties

Physicochemical parameters were calculated using the ExPASy ProtParam tool, including the theoretical isoelectric point (pI), molecular weight, total number of positively and negatively charged amino acids, extinction coefficient (EC), instability index (II), aliphatic index (AI), and the grand average of hydropathicity (GRAVY).:

$$change\ (\%) = \frac{fusion\ protein\ value - wild\ type\ protein\ value}{|wild\ type\ protein\ value|}$$

2.1.1 Prediction of signal peptides and disordered protein regions

The SignalP 6.0 program was used to identify signal peptides in the proteins. Only Sec/signal peptides (Sec/SPI) occurring in eukaryotes were searched. The DeepLoc 2.0 multi-label predictor was used to estimate the intracellular localization of proteins; this tool uses a transformer-based protein language model, providing accurate predictions and interpretability. Only probability values exceeding the DeepLoc 2.0 default thresholds were considered: cytoplasm (0.4761), nucleus (0.5014), and endoplasmic reticulum (0.6090). To validate the predictions, the results were compared with data from the UniProt database and The Human Protein Atlas. Additionally, the NetGPI-1.1 GPI-anchored predictor was used to investigate the glycosylphosphatidylinositol (GPI) anchorage of the proteins.

To predict intrinsically disordered regions (IDRs) in the proteins, IUPred3 was used, which is an online tool based on a biophysical model that identifies regions lacking a stable structure under native conditions. Using the IUPred3 web server, a comparative analysis was performed on the

wild-type BCOR (vBCOR) and its fusion proteins, focusing on the 1448-1633 (region$^{ANK+linker}$) and 1634-1748 (region^{PUFD}) amino acid sections

1.2.3 Analysis of protein structures

Three-dimensional (3D) models of the proteins involved in the BCOR-PCGF1 and PRC1.1 complexes were generated using AlphaFold3 (AF3). Dimers of BCOR and PCGF1, as well as tetramers of BCOR, PCGF1, KDM2B, and S-phase kinase-associated protein 1 (SKP1), were constructed. For each complex, 50 models were created. The generated models were compared with the experimentally determined BCOR-PCGF1 dimer structure (PDB ID: 4hpl) and with models of the wild-type proteins. Due to the sequence length limitations of the AlphaFold Server (beta), we were unable to generate the structure of the 5480-amino-acid-long KMT2D::BCOR fusion.

The PRODIGY (PROtein binDIng enerGY prediction) web server was used to estimate binding affinity, which is described by the Gibbs free energy change (ΔG, kcal mol^{-1}). The PRODIGY web server estimates the ΔG change during binding based on the 3D structure of protein-protein complexes by combining structural and energetic parameters. We investigated the ΔG changes of BCOR-PCGF1 dimers along the full protein sequence (full-length), and specifically for the BCOR PUFD region (1634-1748) and the PCGF1 RAWUL region (167-255; RAWUL-PUFD domain length). Additionally, we analyzed the RAWUL domain regions 167-177 and 185-254, and the PUFD domain region 1636–1748 (4HPL length). For protein structure visualization, PyMOL software was used. The contact map was created with the MAPIYA web server using a distance threshold of 5.5 Å.

1.2.4 Molecular dynamics simulation

To investigate the dimer structures, 10 ns long molecular dynamics (MD) simulations were performed using the GROMACS software package. For each of the nine BCOR-PCGF1 dimers (one wild-type and eight gene fusions), a single structure was selected from the 50 replicates generated by AlphaFold, based on the highest ipTM (interface predicted template modelling) and pTM (predicted template modelling) scores, ensuring that the most stable and reliable conformations were analyzed. The binding affinities between the BCOR proteins and PCGF1 were calculated using the gmx_MMPBSA program, which combines molecular mechanics energies with Poisson-Boltzmann and surface area continuum solvation (MM-PBSA) methods to estimate interaction energies with high accuracy.

1.3 Statistical analysis

Statistical analyses were performed using GraphPad Prism 8.0.1 and 9.5.1 software (GraphPad Software, San Diego, CA, USA). The choice of statistical tests was in every case based on careful consideration of the data type (continuous, categorical), the number of groups to be compared (two or more), the independence or relatedness of the groups, and the data distribution. Before each comparative statistical analysis, we examined the data distribution and the homogeneity of variances. We used the built-in diagnostic tools of GraphPad Prism, including normality tests (e.g., Shapiro-Wilk, Anderson-Darling, D'Agostino & Pearson, Kolmogorov-Smirnov tests), as well as diagnostics of residuals. The software allows for specifying whether to assume a Gaussian distribution and equal standard deviations. If the assumptions of normality and/or equality of variances were not met, as supported by the software's diagnostics, the

appropriate non-parametric tests or more robust versions of ANOVA (e.g., Brown-Forsythe ANOVA) were chosen.

To examine whether there is a difference in the pathogenicity scores for the TP53 protein (scores from various databases and prediction tools, stability values) among the three independent clinical groups (AML-MR, MDS-IB, MDS-LB), an ordinary one-way analysis of variance (ANOVA) was used if the data showed a normal distribution and homogeneous variance within groups. In case of a significant ANOVA result, Tukey's multiple comparisons test was used to identify which specific group means differed significantly from each other. If the assumption of homogeneity of variances was violated, the Brown-Forsythe ANOVA test was applied, followed by Dunnett's T3 multiple comparisons test, as the latter does not assume equal variances. If the data distribution significantly deviated from normality, the non-parametric Kruskal-Wallis test was chosen, which, in case of a significant result, was followed by Dunn's multiple comparisons test for pairwise group comparisons, correcting for the error arising from multiple testing.

A correlation matrix was created using the Pearson correlation coefficient (r) to determine the linear relationship between the continuous variables generated by the stability prediction methods (I-Mutant2.0, DynaMut2, DDGun, and DDGun3D). Before applying the Pearson correlation, we confirmed at least an approximately linear relationship and normality between the variables.

For comparing the IUPred3 scores of IDRs, where multiple, related (non-independent) measurements on the same proteins were compared, the non-parametric Friedman test (the non-parametric equivalent of ANOVA for related samples) was used. In case of a significant result,

Dunn's multiple comparisons test was used to identify pairwise differences.

For comparing the binding affinity values of the BCOR-PCGF1 and PRC1.1 complexes between the different fusion proteins and the wild-type control groups, based on the principles detailed above and following a preliminary examination of the data (normality, equality of variances), we used either ordinary one-way ANOVA (with Tukey's post-hoc test), Brown-Forsythe ANOVA (with Dunnett's T3 post-hoc test), or the Kruskal-Wallis test (with Dunn's post-hoc test). The choice was in every case based on the results of the diagnostic tests performed by GraphPad Prism.

A P value < 0.05 was considered statistically significant for all tests performed. The graphs and figures were created using GraphPad Prism 9.5.1 and the IBS (Illustrator of Biological Sequences) software.

2. Results

2.1 Molecular and *in silico* protein analysis of TP53 in myelodysplastic neoplasias and acute myeloid leukemia

2.1.1 Clinicopathological characteristics

The average age of the 77 patients included in the study was 64.1 years, with a male-to-female ratio of 42:35. Based on their diagnosis, the patients were divided into three groups: 26 patients with acute myeloid leukemia with myelodysplasia-related changes (AML-MR), 12 patients with myelodysplastic syndrome with increased blasts (MDS-IB), and 39 patients with myelodysplastic syndrome with low blasts (MDS-LB).

2.1.2 Next-generation sequencing

Out of the 77 cases studied, at least one TP53 mutation was identified in 26 (33.8%), detecting a total of 41 separate mutations, which constituted 30 different genotypes. The prevalence of mutations was highest in the AML-MR group (57.7%), followed by the MDS-IB (33.3%) and MDS-LB (17.9%) groups. In seven cases, multiple mutations were present within a single sample. The average variant allele frequency (VAF) was highest in the MDS-IB and AML-MR groups. The vast majority (83.3%) of the 30 unique mutations were in the protein's DNA-binding domain (DBD). Most of the mutations (77%) were missense variants, but we also identified frameshift and stop-codon-introducing alterations, several of which caused significant shortening of the protein, a so-called truncation.

2.1.2 Investigation of the relationship between cytogenetic, IHC results, survival, and TP53 mutational status

Out of all *TP53* mutant cases, 11/26 (42.3%) had a complex karyotype (CK). Of the 15 TP53 mutant AML-MR patients, 9 were CK (60%). In the MDS-IB group, 4/2 (50%) cases were proven CK, while in the MDS-LB mutant positive group, no CK was detected, respectively (cytogenetic analysis was not performed in 12 cases). In the AML-MR group, 15/26 (57.7%) cytogenetic aberrations were detected, while in the MDS groups the chromosome alterations were proven with lower frequencies (4/12-33.3% in MDS-IB and 7/38-18.4% in MDS-LB patients, respectively).

IHC staining was positive in a total of 13 cases (16.9%), with 10 (38.5%) positive in the AML-MR group, two (16.7%) in MDS-IB, and 1 (2.6%) in MDS-LB. In total, 14.3% of the cases were IHC and NGS positive, while 42% of all NGS mutants were IHC positive. Nine of 15 NGS mutants were IHC positive in the AML-MR group (60%), 50% in the MDS-IB group, and no IHC and NGS double-positive cases were found in the MDS-LB patients. Of the six samples that resulted in truncated proteins, five had negative staining results following IHC. The p.R213X, p.Y163Xfs, p.E286Qfs, p.L93Lfs as single mutations, p.C135X with another two mutations within one sample, were IHC negative, while p.Y220X in parallel with another mutation was IHC positive.

A significant difference was found in the median OS between the 3 groups ($P \leq 0.0001$) with respect of mutant/wild *TP53* status (918 days for wild-type cases-AML-MR: 341, MDS-IB: 260, MDS-LB: 1101, while 224 days for mutant patients-AML-MR: 199, MDS-IB: 58.5, MDS-LB: 626 days, as well).

2.1.3 *In silico* prediction of the functional effects of mutations

A significant portion of the identified missense mutations were classified by the ClinVar database as being of uncertain clinical significance or conflicting, which justified further detailed bioinformatic analysis. We evaluated the potential deleterious effects of the mutations using several programs that predict pathogenicity and protein stability based on different principles. The results indicated that most of the studied mutations likely had pathogenic functional consequences. *In vitro* experimental data available in the IARC TP53 database confirmed that mutations located in the DNA-binding domain exhibited, on average, only 13.8% of the transcriptional activity of the wild-type protein. A significant portion of these mutations had a loss-of-function effect, while a smaller part had a dominant-negative or even a gain-of-function effect.

2.1.4 Analysis of the stability and interactions of mutant TP53 proteins

In silico stability studies showed that most of the analyzed mutations significantly reduced the structural stability of the TP53 protein. A strong correlation was observed between the results of the different prediction methods, which increased the reliability of the predictions. Protein-protein interaction modeling revealed that the p.G334R mutation weakened the binding between TP53 monomers, which could inhibit the formation of the tetramer structure essential for function. The p.S362N variant, in turn, altered the strength of the interaction between TP53 and its regulatory protein, USP7/HAUSP.

A comparative analysis of the AML-MR, MDS-IB, and MDS-LB groups revealed that the mutations identified in the AML-MR group had the most severe functional consequences in terms of pathogenicity scores, protein stability reduction, and loss of transcriptional activity. Based on these indicators, the MDS-IB group represented an intermediate state between the AML-MR and the lower-risk MDS-LB groups.

2.2 *In silico* analysis of *BCOR*-rearranged sarcomas

2.2.1 Sequence and domain characteristics of the fusion proteins

To investigate the pathomechanism of rare BCOR-rearranged sarcomas (BRS), we analyzed nine known gene fusions described in the literature using comprehensive bioinformatic methods. The analysis revealed that although the fusion events caused significant structural changes, the crucial PCGF1-binding PUFD domain of the BCOR protein was retained in all examined fusion proteins. However, other functional domains were often lost during the fusions, such as the BCL6-binding domain (Bbs) of BCOR, which could lead to the impairment of the protein's transcriptional corepressor function. The fusion partners also underwent significant truncation, losing critical functional regions, such as the destruction box of CCNB3 and the SET domain of KMT2D, which is responsible for histone methyltransferase activity.

2.2.2 Physicochemical and localization properties of the fusion proteins

The *in silico* analysis of the fusion proteins revealed a marked alteration in their physicochemical properties compared to the wild-type proteins. In general, the instability index of the fusion proteins increased, while their hydrophobicity decreased, indicating a deterioration of the

proteins' structural stability and potentially altered solubility. Localization prediction models predicted nuclear localization for most fusion proteins-similar to wild-type BCOR. The analysis of the proteins' IDRs revealed a significantly increased structural disorder in the C-terminal region of the PUFD domain in some of the fusion proteins, which could fundamentally affect the protein's binding capabilities and the proper assembly of the PRC1.1 complex.

2.2.3 Three-dimensional modeling of the complexes and static analysis of binding affinity

To gain a deeper understanding of the functional consequences of the fusion events, we modeled the three-dimensional structures of the complexes formed by the fusion proteins described in BRS, the wild-type BCOR (vBCOR), and their partner PCGF1. For the modeling, we used the AF3 deep learning algorithm, which can predict the spatial structures of proteins and their complexes with high accuracy. The modeling extended to BCOR-PCGF1 dimers as well as the broader tetramer complexes containing subunits of the non-canonical Polycomb repressive complex 1 (PRC1.1), namely KDM2B and SKP1.

To quantify the stability of the complexes and the strength of the interactions between the proteins, we used the PRODIGY web server, which estimates binding affinity based on the 3D structure, expressed as the Gibbs free energy change (ΔG). The analyses revealed significant changes in binding affinity between the wild-type and fusion complexes. Based on the results, the fusion proteins could be divided into two groups: some fusions, such as BCOR::CCNB3, showed increased binding affinity (lower ΔG value), suggesting the formation of an overly stable, rigid complex; whereas other fusions, such as ZC3H7B::BCOR and

CIITA::BCOR, had significantly reduced affinity (higher ΔG value), indicating complex instability and weaker binding. The proper epigenetic function of the PRC1.1 complex requires a delicate balance of stability and dynamic flexibility for the assembly and disassembly processes essential for gene regulation. According to our results, both excessive stabilization and weakened binding disrupt this delicate dynamic, which can lead to loss of complex function.

2.2.4 Molecular dynamics analysis of the dimer complexes

Following the analysis of the static 3D models, we performed 10-nanosecond MD simulations with the GROMACS software package to gain dynamic insight into the behavior of the BCOR-PCGF1 dimer complexes. The MD simulations confirmed and further refined the conclusions drawn from the static models.

The root-mean-square deviation analysis showed that the fusion complexes generally exhibited greater structural instability and conformational flexibility compared to the wild-type complex, which remained more stable during the simulation. The ZC3H7B::BCOR and CIITA::BCOR fusions showed high fluctuating RMSD values, indicating sustained structural instability.

Analysis of the number of hydrogen bonds between the two proteins revealed that fusion complexes with weaker binding affinity (e.g., ZC3H7B::BCOR) formed significantly fewer stable hydrogen bonds with the PCGF1 protein, which directly explains the reduced binding force and complex instability.

The investigation of the radius of gyration showed that the fusion complexes were less compact and adopted a more extended conformation than the tightly fitting wild-type dimer. This structural "loosening"

suggests that the fusion events alter the overall spatial arrangement of the complex, which may affect its function.

The binding energies calculated from the MD simulations showed a strong correlation with the static ΔG values estimated by PRODIGY, which supports the reliability of the results and provides a consistent picture of the thermodynamic changes caused by the fusions.

Overall, these analyses provide a mechanistic explanation for the paradox of why the function of the PRC1.1 complex is impaired in BRS despite the retention of the crucial PUFD domain. The fusions contribute to errors in gene regulation and the development of tumors not only through the loss of domains but also by altering the fine structural and dynamic properties of the proteins-by increasing flexibility, weakening interactions, and destabilizing the global structure

.

3. Main Findings and Conclusion

Briefly summarized, the main findings of this Ph.D. work are as follows:

- Successfully performed molecular and *in silico* analysis of TP53 mutations in 77 patients with myelodysplastic syndrome (MDS) and acute myeloid leukemia (AML), identifying a total of 41 *TP53* mutations, including 30 unique types, in 26 patients (33.8%).
- The frequency of *TP53* mutations was significantly higher in the AML-MR (AML with myelodysplasia-related changes) subgroup (57.7%) compared to the MDS-IB (MDS with increased blasts, 33.3%) or MDS-LB (MDS with low blasts, 17.9%) subgroups, and these mutations were associated with poorer survival and complex karyotype.
- *In silico* analyses predicting pathogenicity, stability, and transcriptional activity revealed significant differences among *TP53* mutations detected in AML-MR, MDS-IB, and MDS-LB groups, supporting the different clinical courses and prognoses.
- *In silico* modeling of protein-protein (e.g., homotetramerization, USP7 binding) and protein-DNA interactions highlighted the functional consequences of specific *TP53* mutations, such as impaired tetramerization or weakened DNA binding, contributing to the understanding of their pathogenic mechanisms.
- The applied *in silico* tools (e.g., Align-GVGD, REVEL, BayesDel, I-Mutant2.0, DynaMut2, DDGun, mCSM-PPI2, mCSM-NA) proved useful for evaluating variants of uncertain significance (VUS) in the TP53 gene and predicting their pathogenicity in MDS and AML.

- Performed a comprehensive *in silico* functional analysis of nine previously described BCOR gene fusions occurring in BCOR-rearranged sarcoma (BRS), examining sequential, domain structural, physicochemical, localization, and disorder properties.
- Established that although the PUFD domain of the BCOR protein, critical for binding to the PRC1.1 complex, was retained in all fusion proteins studied, the fusions induced significant structural and physicochemical changes compared to the wild-type proteins.
- 3D structure modeling and binding affinity calculations using AlphaFold3 and PRODIGY showed that BCOR fusions significantly affect the dimerization between BCOR and the PCGF1 protein; some fusions (e.g., BCOR::CCNB3, BCOR::MAML3) increased, while others (e.g., ZC3H7B::BCOR, CIITA::BCOR) decreased the binding affinity compared to the wild-type complex.
- Modeling and binding affinity analysis of the non-canonical PRC1.1 complex (BCOR/Fusion, PCGF1, KDM2B, SKP1) suggest that BCOR fusion proteins may disrupt the normal assembly and/or epigenetic function of the complex.
- *In silico* modeling proved to be a valuable toolkit for investigating the pathomechanism of BCOR fusions underlying BRS, enabling detailed analysis of the structural and interaction changes caused by the fusion proteins.

4. Publication List

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/164/2025.PL
Subject: PhD Publication List

Candidate: Kristóf Madarász
Doctoral School: Doctoral School of Clinical Medicine
MTMT ID: 10088102

List of publications related to the dissertation

1. **Madarász, K.**, Mótyán, J. A., Chang Chien, Y. C., Bedekovics, J., Csoma, S. L., Méhes, G., Mokánszki, A.: BCOR-rearranged sarcomas: In silico insights into altered domains and BCOR interactions.
   Comput. Biol. Med. 191, 1-22, 2025.
   DOI: http://dx.doi.org/10.1016/j.compbiomed.2025.110144
   IF: 7 (2023)
2. **Madarász, K.**, Mótyán, J. A., Bedekovics, J., Miltényi, Z., Ujfalusi, A., Méhes, G., Mokánszki, A.: Deep Molecular and In Silico Protein Analysis of p53 Alteration in Myelodysplastic Neoplasia and Acute Myeloid Leukemia.
   Cells. 11 (21), 1-23, 2022.
   DOI: http://dx.doi.org/10.3390/cells11213475
   IF: 6

List of other publications

3. Bedekovics, J., **Madarász, K.**, Mokánszki, A., Deliné Molnár, S., Mester, Á., Miltényi, Z., Méhes, G.: Exploring p53 protein expression and its link to TP53 mutation in myelodysplasia-related malignancies - Interpretive challenges and potential field of applications.
   Histopathology. 85 (1), 143-154, 2024.
   DOI: http://dx.doi.org/10.1111/his.15185
   IF: 3.9 (2023)
4. Csoma, S. L., **Madarász, K.**, Chang Chien, Y. C., Emri, G., Bedekovics, J., Méhes, G., Mokánszki, A.: Correlation Analyses between Histological Staging and Molecular Alterations in Tumor-Derived and Cell-Free DNA of Early-Stage Primary Cutaneous Melanoma.
   Cancers (Basel). 15 (21), 1-13, 2023.
   DOI: http://dx.doi.org/10.3390/cancers15215141
   IF: 4.5

UNIVERSITY of DEBRECEN

UNIVERSITY AND NATIONAL LIBRARY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

5. Chang Chien, Y. C., **Madarász, K.**, Csoma, S. L., Mótyán, J. A., Huang, H. Y., Méhes, G., Mokánszki, A.: Molecular Identification and In Silico Protein Analysis of a Novel BCOR-CLGN Gene Fusion in Intrathoracic BCOR-Rearranged Sarcoma.
Cancers (Basel). 15 (3), 1-17, 2023.
DOI: http://dx.doi.org/10.3390/cancers15030898
IF: 4.5

Total IF of journals (all publications): 25,9
Total IF of journals (publications related to the dissertation): 13

The Candidate's publication data submitted to the Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

28 May, 2025

Acknowledgements

I want to thank my supervisor, Dr. Attila Mokánszki, Assistant Professor at the Institute of Pathology, Faculty of Medicine, University of Debrecen, for his support and patience during my research. This work would not have been possible without his professional advice and scientific precision.

I am deeply grateful to the director of the Institute of Pathology, Prof. Dr. Gábor Méhes, who provided the opportunity to conduct my research and supported me over the past years.

Sincere thanks are due to the staff of the Molecular Laboratory at the Institute of Pathology, who not only introduced me to the world of routine diagnostics but also significantly enriched my research work with their friendship and support. I would especially like to thank Anikó Mónus and Szilvia Csoma for their selfless help and professional support over the many years.

I owe my gratitude to Dr. János András Mótyán, Associate Professor at the Institute of Biochemistry and Molecular Biology, for his valuable help during my work and for introducing me to the world of applying *in silico* methods, thereby significantly contributing to my professional development.

Our research was supported by the following grants: ÚNKP-23-3-II New National Excellence Program Doctoral Student Research Scholarship, GTIDEA Excellence PhD Scholarship, University of Debrecen Doctoral Student Supplementary Scholarship, EKÖP-24-3-II Doctoral Student Scholarship II.