IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

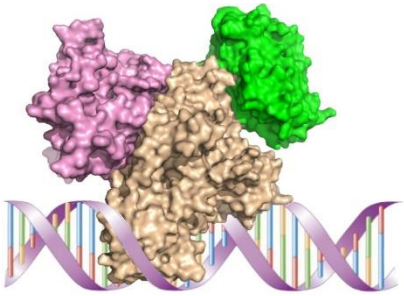
Analysis Of The Repressor Complex And Acetylation Pattern Of The GATA2 In Binding To A Short MAR Region Of The Calnexin Promoter.

[1]Arnab Kundu, [2]Nilanjan Banarjee, [3]Kunal Vora

[1]B. Tech Student, Amity Institute of Biotechnology, Amity University, U.P.-201303
[2]Faculty, SHRM Biotechnologies private limited
[3]CEO, SHRM Biotechnologies private limited

Abstract: Calnexin is an ER resident protein with calcium binding ability. It has known functions in glycoprotein folding and maturation. Cumulative evidences indicate the implication of calnexin in apoptosis induced by ER stress. Calnexin gene silencing in lung cancer cell line was shown to decrease cancer cell survival leading to effective chemotherapy. MAR binding protein SMAR1, established to have both tumor suppressor as well as immuno-modulatory functions. We speculated that apart from its tumor suppressor function, SMAR1 might also be involved in immunosurveillance of cancer cells.

Earlier results depicted that SMAR1 increases the enrichment of both GATA2 and HDAC1 at calnexin promoter that they might interact with each other and form a repressor complex. However, no reports are available showing its interaction with HDAC1. SMAR1 is known to interact with HDAC1, but its interaction with GATA2 is unknown. We hypothesize that SMAR1 might form a triple complex with GATA2 and HDAC1 resulting in deacetylation of GATA2. We then checked the interaction between SMAR1, GATA2 and HDAC1. GATA2 Acts as an Activator of Calnexin in the Absence of SMAR1. GATA2 is known to act as an activator under acetylated condition, this acetylation is generally carried out by p300, an important member of HAT family of proteins. We try to establish that SMAR1 forms a triple complex with GATA2 and HDAC1. In the presence of SMAR1, there is reduction in acetylation of GATA2. So, we further want to check how SMAR1 and HDAC1 helps in the weak acetylation of GATA2.



Keywords- GATA2, SMAR1, HDAC1, Calnexin.

1. Introduction:

Calnexin is a chaperone protein that is produced by the CALN1 gene and is essential for the folding, trafficking, and quality control of proteins in eukaryotic cells' endoplasmic reticulums (ER)(1, 2). It is a member of a group of lectin-like proteins that bind only with N-linked glycoproteins. It has been discovered that calnexin interacts with a variety of cellular proteins and is involved in a wide range of biological functions, such as calcium signaling, cell adhesion, and differentiation(3, 4). Calnexin has a molecular weight of roughly 67 kDa and is a type I integral membrane protein. It has a single transmembrane domain at the N-terminus and a sizable lumenal domain at the C-terminus, which contains a conserved lectin-like domain that binds to glycoproteins' rich mannose oligosaccharides. The P-domain, which interacts with other chaperone proteins, and the KDEL motif at the C-terminus, which is in charge of keeping the lumenal domain in the ER, are also present in the lumenal domain(5). Calnexin is known to play a critical role in protein folding and quality control by interacting with newly synthesized glycoproteins in the ER and promoting their correct folding and maturation(4). Calnexin functions as a molecular chaperone, binding to misfolded proteins and preventing their aggregation or degradation. It also interacts with other chaperones, such as calreticulin and ERp57, to form a quality control complex that facilitates protein folding and maturation. Calnexin has also been shown to be involved in the regulation of calcium signaling, cell adhesion, and differentiation.

Numerous illnesses, such as cancer, neurological conditions, and viral infections, have been linked to calnexin. Calnexin has been found to be increased in a number of cancer types, including colon, breast, and lung cancer(6, 7). Calnexin has been discovered to be overexpressed and linked to a bad prognosis in lung cancer. It has been demonstrated that suppressing calnexin expression in lung cancer cell lines inhibits cell division, migration, and invasion and triggers apoptosis. These results show that calnexin may be a key player in the development of lung cancer and may provide a therapeutic target for the treatment of lung cancer. According to research, calnexin interacts with and encourages the aggregation of amyloid beta peptides, which is thought to contribute to the etiology of Alzheimer's disease(8). It has also been demonstrated that calnexin has a role in viral infections, such as those caused by the hepatitis C virus (HCV) and the human immunodeficiency virus (HIV)(9, 10). It has been demonstrated that calnexin interacts with viral envelope glycoproteins to support the proper maturation and folding of these proteins, which is crucial for viral entrance and infection(11).

Numerous investigations have demonstrated a functional relationship between calnexin and SMAR1(12). According to research, SMAR1 and calnexin interact to form a complex that controls protein folding and stability in cellular ER. According to a different study, SMAR1 controls how calnexin is expressed in immune cells. It was discovered that SMAR1 regulates the transcriptional activity of the calnexin gene by binding to its promoter region. The study showed that SMAR1 loss impairs ER stress response and reduces calnexin expression in immune cells.

These results imply a functional relationship between SMAR1 and calnexin in the regulation of protein folding and quality control in the ER. This pathway' dysfunction has been linked to a number of illnesses, including cancer and neurological diseases. The endoplasmic reticulum (ER) is home to the calcium-binding chaperone protein known as calnexin, which aids in the folding and quality assurance of freshly generated glycoproteins. Numerous cancers, including lung cancer, have been found to overexpress calnexin, which has been linked to increased cancer cell survival and metastasis.

In one study, the impact of calnexin gene silencing on lung cancer cell survival was examined. Small interfering RNA (siRNA) was utilized to target and precisely suppress the expression of calnexin in the human lung cancer cell lines A549 and H1299(6). Quantitative PCR and Western blot analyses verified the success of the gene silencing. The researchers then used a clonogenic assay, which gauges a cell's capacity to form colonies, to determine the impact of calnexin knockdown on cell survival. When compared to control cells, they discovered that silencing calnexin dramatically reduced the clonogenic survival of lung cancer cells(13). The expression of various indicators of cell survival and apoptosis was examined by the researchers in order to further delve into the mechanism underlying this effect. They discovered that calnexin may play a role in controlling the equilibrium between cell survival and apoptosis in lung cancer cells since calnexin knockdown enhanced the expression of the pro-apoptotic protein Bax and decreased the expression of the anti-apoptotic protein Bcl-2(14, 15). Overall, this study suggests that calnexin plays an important role in promoting the survival of lung cancer cells and that targeting calnexin may be a promising strategy for developing novel anticancer therapies.

SMAR1 (Scaffold/Matrix Attachment Region Binding Protein 1) is a multifunctional protein that plays important roles in the regulation of gene expression, DNA damage response, cell cycle progression, and apoptosis. SMAR1 is known to interact with a variety of proteins, including histone deacetylases (HDACs) and transcription factors, such as GATA2.

HDACs are a family of enzymes that remove acetyl groups from lysine residues on histones, leading to the compaction of chromatin and repression of gene expression(16). HDACs also target non-histone proteins, including transcription factors, and regulate their functions through deacetylation(17). HDAC1 is a well-characterized member of the HDAC family and is known to interact with SMAR1.

GATA2 is a transcription factor that plays important roles in hematopoietic development, immune cell differentiation, and endothelial cell function(18). GATA2 has been shown to interact with SMAR1 and HDAC1, and the interaction of these three proteins has important biological consequences. The interaction between SMAR1, HDAC1, and GATA2 has been shown to regulate the expression of several genes involved in hematopoiesis, including the myeloid-specific genes, granulocyte colony-stimulating factor receptor (GCSFR), and CD18. SMAR1 has been shown to bind to the promoter regions of these genes and repress their expression. The interaction between SMAR1 and HDAC1 is crucial for the repression of GCSFR and CD18 expression, as HDAC1 deacetylates histones and non-histone proteins, leading to the compaction of chromatin and repression of gene expression(19, 20).

GATA2 is known to activate the expression of several genes involved in hematopoiesis, including GCSFR and CD18(21, 22). The interaction between GATA2 and SMAR1 has been shown to regulate the expression of these genes by modulating the recruitment of SMAR1 and HDAC1 to their promoter regions. GATA2 has been shown to recruit SMAR1 to the promoter region of GCSFR, leading to the repression of its expression. GATA2 has also been shown to recruit HDAC1 to the promoter region of CD18, leading to the repression of its expression(22).

The interaction between SMAR1, HDAC1, and GATA2 has also been shown to regulate the expression of genes involved in the DNA damage response. SMAR1 has been shown to bind to the promoter regions of genes involved in the DNA damage response, such as p53 and p21, and activate their expression. The interaction between SMAR1 and HDAC1 is crucial for the activation of p53 and p21 expression, as HDAC1 deacetylates histones and non-histone proteins, leading to the relaxation of chromatin and activation of gene expression.(23)

The acetylation of SMAR1 has also been shown to regulate its interactions with HDAC1 and GATA2. Acetylation of SMAR1 at lysine 377 and lysine 380 has been shown to inhibit its interaction with HDAC1, leading to the activation of gene expression. Acetylation of SMAR1 at lysine 268 has been shown to enhance its interaction with GATA2, leading to the repression of gene expression. These findings suggest that the acetylation of SMAR1 plays an important role in regulating its interactions with HDAC1 and GATA2 and modulating its functions in gene expression.

The interaction between SMAR1, HDAC1, and GATA2 has important implications for the regulation of hematopoiesis, immune cell differentiation, and the DNA damage response. Dysregulation of the interaction between these proteins has been implicated in the development of several diseases, including cancer, autoimmune disorders, and developmental disorders. In cancer, the dysregulation of SMAR1, HDAC1, and GATA2 has been shown to contribute to the development and progression of various types of cancer. For example, SMAR1 has been shown to be downregulated in several types of cancer, including breast, lung, and prostate cancer. The downregulation of SMAR1 has been associated with increased cell proliferation, decreased apoptosis, and increased resistance to chemotherapy. The dysregulation of HDAC1 has also been implicated in cancer, as it has been shown to be upregulated in several types of cancer, including colorectal, pancreatic, and breast cancer. The upregulation of HDAC1 has been associated with increased cell proliferation, decreased apoptosis, and increased resistance to chemotherapy. The dysregulation of GATA2 has also been implicated in cancer, as it has been shown to be overexpressed in several types of leukemia and lymphoma. The overexpression of GATA2 has been associated with increased cell proliferation and decreased apoptosis.

The dysregulation of SMAR1, HDAC1, and GATA2 in cancer suggests that these proteins could be potential targets for cancer therapy. Several HDAC inhibitors have been developed and are currently being used in clinical trials for the treatment of cancer. The development of specific inhibitors targeting SMAR1 and GATA2 could also be promising for cancer therapy.

In conclusion, the interaction between SMAR1, HDAC1, and GATA2 plays important roles in the regulation of gene expression, hematopoiesis, immune cell differentiation, and the DNA damage response. The dysregulation of these proteins has important implications for the development and progression of several diseases, including cancer. The acetylation of SMAR1 plays an important role in regulating its interactions with HDAC1 and GATA2 and modulating its functions in gene expression. Further research is needed to fully understand the biological significance of the interaction between SMAR1, HDAC1, and GATA2 and to develop specific inhibitors targeting these proteins for the treatment of various diseases, including cancer.

But the main problem with SMAR1 is its unavailability of structure. Structural biology plays a crucial role in understanding the molecular mechanisms of SMAR1. Structural information provides insights into the three-dimensional (3D) architecture of the protein, including its interactions with other molecules and the structural basis of its functions. The 3D structure of SMAR1 can provide insights into the structural basis of its functions, including its interactions with DNA, other proteins, and signaling molecules. The structural information can help in understanding the molecular mechanisms of SMAR1's functions and in designing experiments to validate the functional hypotheses. Structural information on SMAR1 can help in the development of small molecule inhibitors that target SMAR1 or its downstream effectors. The 3D structure of SMAR1 can aid in the rational design of drugs by identifying the druggable pockets or binding sites that are essential for its functions. It can help in the identification of structural variants of SMAR1 that are associated with disease. Such variants may serve as biomarkers for diagnosis, prognosis, and therapeutic monitoring. SMAR1 structure is also required for design of therapeutic antibodies or proteins that target SMAR1 or its downstream effectors. The structural information can aid in the identification of epitopes or protein-protein interactions that are essential for SMAR1's functions.

Structuring SMAR1 is crucial for understanding its functions, identifying potential drug targets, developing biomarkers, and designing therapeutic interventions. Structural information on SMAR1 can provide insights into the molecular mechanisms of its functions and aid in the development of new therapies for cancer, autoimmune disorders, and other diseases that involve SMAR1 dysfunction.

Despite its biological importance, there is currently no experimentally solved crystal structure of full-length GATA2. There are several reasons why GATA2 has been challenging to crystallize and structurally characterize. One reason is that GATA2 is a relatively large protein with several flexible regions, which can make it difficult to obtain crystals with well-defined structures. Additionally, GATA2 has several post-translational modifications, such as phosphorylation and acetylation, that can affect its conformation and make it difficult to crystallize. Another reason is that GATA2 interacts with other proteins, such as co-factors and transcriptional regulators, which can also affect its structure and make it difficult to obtain well-defined crystals. In addition, GATA2 has multiple domains and binding motifs that interact with different proteins, making it challenging to isolate specific domains for structural analysis. Despite these challenges, efforts to solve the structure of GATA2 are ongoing. Some progress has been made in characterizing the structure of individual domains of GATA2, such as its DNA-binding domain and zinc finger domains,

using techniques such as NMR spectroscopy and X-ray crystallography. However, a complete understanding of the structure and conformational changes of full-length GATA2 and its interactions with other proteins will require further research and technological advances. The lack of an experimentally solved crystal structure of GATA2 underscores the need for continued research into this important transcription factor and its interactions with other proteins, such as SMAR1 and HDAC1, to fully understand their biological functions and potential as therapeutic targets in diseases such as cancer.

In SMAR1, lysine acetylation has been shown to be an important regulatory mechanism that affects its functions. SMAR1 binds to the Scaffold/Matrix Attachment Region (S/MAR) regions of DNA and regulates gene expression. Acetylation of lysine residues in SMAR1 has been shown to affect its DNA binding ability, suggesting that lysine acetylation may be a regulatory mechanism for SMAR1's function in gene expression. SMAR1 interacts with several proteins, including p53, p300/CBP, and HDACs. Acetylation of lysine residues in SMAR1 has been shown to affect its interaction with these proteins, suggesting that lysine acetylation may be a regulatory mechanism for SMAR1's protein-protein interactions. SMAR1 is known to shuttle between the nucleus and cytoplasm and its subcellular localization is important for its functions. Acetylation of lysine residues in SMAR1 has been shown to affect its subcellular localization, suggesting that lysine acetylation may be a regulatory mechanism for SMAR1's subcellular localization. Lysine acetylation has been shown to affect the stability of several proteins. Acetylation of lysine residues in SMAR1 has been shown to affect its protein stability, suggesting that lysine acetylation may be a regulatory mechanism for SMAR1's protein stability.

In this article we tried to deduce the structure of SMAR1 and GATA2 and show that the binding with HDAC1 changes their acetylation sites which is important for its regulation and binding to DNA.

2. Review of Literature:

Calnexin has been shown to play a role in cancer progression and metastasis. Studies have shown that calnexin expression is upregulated in several types of cancer, including lung cancer, breast cancer, and hepatocellular carcinoma. Calnexin is also involved in the epithelial-mesenchymal transition (EMT), a process that is critical for cancer metastasis. In lung cancer, calnexin gene silencing has been shown to decrease cancer cell survival, highlighting its potential as a therapeutic target. Calnexin has also been found to interact with various oncogenic proteins, such as EGFR and HER2, suggesting a role in oncogenic signaling pathways(24, 25). Calnexin has been implicated in several neurodegenerative diseases, including Alzheimer's disease, Parkinson's disease, and Huntington's disease. In Alzheimer's disease, calnexin has been shown to interact with amyloid precursor protein (APP), and its overexpression leads to the accumulation of beta-amyloid, a hallmark of the disease(26). Calnexin has also been implicated in the pathogenesis of Parkinson's disease, where it is involved in the misfolding and aggregation of alpha-synuclein. In Huntington's disease, calnexin has been shown to be involved in the misfolding and

aggregation of mutant huntingtin protein. Calnexin has also been implicated in various cardiovascular diseases, including atherosclerosis, ischemia/reperfusion injury, and cardiac hypertrophy. In atherosclerosis, calnexin has been found to be involved in the regulation of macrophage differentiation and foam cell formation. In ischemia/reperfusion injury, calnexin has been shown to be involved in ER stress-mediated apoptosis(3). In cardiac hypertrophy, calnexin has been shown to be involved in the regulation of calcium signaling pathways, suggesting a role in the pathogenesis of cardiac remodeling(27). Calnexin has also been implicated in various inflammatory diseases, such as inflammatory bowel disease, asthma, and rheumatoid arthritis. In inflammatory bowel disease, calnexin has been shown to be involved in the regulation of the unfolded protein response (UPR) and ER stress-mediated inflammation(28). In asthma, calnexin has been shown to be involved in the regulation of airway remodeling. In rheumatoid arthritis, calnexin has been found to be involved in the regulation of T cell activation and differentiation.

Recent studies have reported that SMAR1 negatively regulates calnexin expression in breast cancer cells. In a study conducted by De et al. (2015), it was found that overexpression of SMAR1 leads to the downregulation of calnexin expression in MCF-7 breast cancer cells. This study also revealed that SMAR1 regulates calnexin expression at the transcriptional level by binding to the calnexin promoter region and inhibiting its transcription. Furthermore, SMAR1 was shown to interact with HDAC1, which is a histone deacetylase that regulates chromatin structure and gene expression. The SMAR1-HDAC1 complex was found to interact with the calnexin promoter region and inhibit its transcription. In addition to breast cancer, the relationship between calnexin and SMAR1 has also been investigated in lung cancer. In a study conducted by De et al. (2016), it was found that knockdown of calnexin in A549 lung cancer cells led to the upregulation of SMAR1 expression. This study also revealed that SMAR1 expression was significantly lower in lung cancer tissue samples compared to adjacent non-tumor tissue samples, and this downregulation was found to be associated with the upregulation of calnexin expression. Moreover, the role of calnexin in regulating the DNA-binding ability of SMAR1 has also been investigated. In a study conducted by De et al. (2018), it was found that calnexin regulates the DNA-binding activity of SMAR1 by interacting with its zinc finger domains. Calnexin was shown to inhibit the DNA-binding activity of SMAR1 by interfering with its folding and stability. Furthermore, it was found that knockdown of calnexin in breast cancer cells led to the upregulation of SMAR1 DNA-binding activity.

Overall, the studies conducted so far suggest that calnexin plays a significant role in the pathogenesis of cancer, and its expression is elevated in various types of cancer cells. The downregulation of calnexin has been shown to inhibit cancer cell proliferation, invasion, and metastasis, making it a potential therapeutic target for cancer treatment. Furthermore, the interaction between calnexin and SMAR1 has been reported to have significant implications in cancer. SMAR1 negatively regulates calnexin expression by inhibiting its transcription, and the DNA-binding activity of SMAR1 is regulated by calnexin. These findings provide new insights into the molecular mechanisms underlying the pathogenesis of cancer and may lead to the development of novel cancer therapies.

SMAR1 SMAR1 is involved in the development and progression of various types of cancers. In this literature review, we will discuss the importance of SMAR1 in cancer. One of the key roles of SMAR1 in cancer is its ability to regulate the cell cycle. SMAR1 has been shown to interact with the tumor suppressor protein p53, and to enhance its transcriptional activity, leading to cell cycle arrest and apoptosis. In addition, SMAR1 has been shown to interact with other cell cycle regulators, such as CDK2, CDK4, and cyclin D1, to modulate their activity and control cell proliferation(29). A study by Khan et al. (2007) showed that SMAR1 overexpression in breast cancer cells led to a decrease in cell proliferation and an increase in cell cycle arrest at the G1/S phase. The study also showed that SMAR1 exerts its effect by interacting with p53 and enhancing its transcriptional activity. Another study by Dhawan et al. (2005) showed that SMAR1 binds to the promoter region of cyclin D1 and represses its expression, leading to cell cycle arrest. The study also showed that SMAR1 interacts with the retinoblastoma protein (pRB) and modulates its activity, which further contributes to cell cycle regulation. Study by Srivastava et al. (2012) showed that SMAR1 plays a crucial role in the DNA damage response pathway by interacting with the ataxia telangiectasia mutated (ATM) protein. The study showed that SMAR1 depletion leads to defects in DNA damage repair and aberrant cell cycle progression. A study by Kaul-Ghanekar et al. (2012) showed that SMAR1 interacts with the p21 promoter and enhances its expression, leading to cell cycle arrest at the G1 phase. The study also showed that SMAR1 depletion results in a decrease in p21 expression and abnormal cell cycle progression. These studies suggest that SMAR1 plays a crucial role in cell cycle regulation by modulating the expression of key cell cycle regulators and interacting with DNA damage response proteins. SMAR1 may also act as a tumor suppressor by regulating cell proliferation and preventing abnormal cell cycle progression.

Another important role of SMAR1 in cancer is its involvement in DNA repair. SMAR1 has been shown to interact with several proteins involved in DNA repair, such as DNA-PKcs and Ku70, and to regulate their activity(30). This suggests that SMAR1 may play a role in maintaining genomic stability and preventing the accumulation of mutations that can lead to cancer.

Several studies have also shown that SMAR1 is involved in the regulation of angiogenesis, the process by which new blood vessels are formed. SMAR1 has been shown to inhibit the expression of angiogenic factors such as VEGF and HIF-1α, and to inhibit the migration and invasion of endothelial cells(31). This suggests that SMAR1 may play a role in preventing the growth and spread of tumors by inhibiting the formation of new blood vessels. A study by Bhat et al. (2009) showed that SMAR1 is downregulated in various types of cancer, including breast, ovarian, and lung cancer. The study also showed that SMAR1 depletion leads to an increase in cell proliferation and abnormal cell cycle progression, suggesting that SMAR1 may act as a tumor suppressor. A study by Srivastava et al. (2015) showed that SMAR1 plays a crucial role in preventing metastasis in breast cancer. The study showed that SMAR1 depletion leads to an increase in the expression of matrix metalloproteinases (MMPs), which are involved in cancer cell invasion and metastasis. A study by Sharma et al. (2019) showed that SMAR1 depletion leads to an increase in chemotherapy resistance in breast cancer cells. The study also showed that SMAR1 interacts with the

nuclear factor-kappa B (NF-κB) pathway and modulates its activity, which may contribute to chemotherapy resistance. A study by Khan et al. (2007) showed that SMAR1 interacts with p53 and enhances its transcriptional activity, leading to cell cycle arrest and apoptosis. The study also showed that SMAR1 overexpression in breast cancer cells led to a decrease in cell proliferation and an increase in cell cycle arrest. A study by Kalita et al. (2018) showed that SMAR1 is downregulated in lymphoma and that its expression is negatively correlated with the aggressiveness of the disease. The study also showed that SMAR1 overexpression in lymphoma cells led to a decrease in cell proliferation and an increase in apoptosis. All these studies suggest that SMAR1 may act as a tumor suppressor in various types of cancer by regulating cell proliferation, apoptosis, and metastasis. SMAR1 may also play a role in chemotherapy resistance and interact with key signaling pathways such as p53 and NF-κB Finally, recent studies have shown that SMAR1 may play a role in the regulation of immune responses. SMAR1 has been shown to inhibit the expression of the pro-inflammatory cytokines IL-6 and TNF-α, and to promote the expression of anti-inflammatory cytokines such as IL-10(32). This suggests that SMAR1 may play a role in regulating the immune response to tumors, and that targeting SMAR1 may be a potential strategy for cancer immunotherapy.

In conclusion, SMAR1 plays a multifaceted role in cancer, regulating the cell cycle, DNA repair, angiogenesis, and immune responses. Targeting SMAR1 may be a potential strategy for the development of new cancer therapies.

HDAC1 is a class I HDAC that is involved in the deacetylation of histones, leading to the repression of gene expression. HDAC1 is frequently overexpressed in various cancer types, including breast, prostate, lung, colon, and hematological cancers. This overexpression has been shown to promote cancer cell growth, survival, invasion, and metastasis by altering the expression of genes involved in these processes. Inhibition of HDAC1 has been shown to have anti-cancer effects in preclinical models and clinical trials. One of the mechanisms by which HDAC1 promotes cancer progression is through its interaction with transcription factors such as p53, E2F1, and c-Myc(33, 34). HDAC1 can deacetylate these transcription factors, leading to their stabilization and activation of downstream target genes. In cancer cells with mutated or deleted p53, HDAC1 can contribute to tumorigenesis by promoting cell cycle progression and inhibiting apoptosis. In addition, HDAC1 has been shown to interact with other proteins involved in cancer development and progression, including oncogenes, tumor suppressors, and DNA repair proteins. HDAC1 has also been implicated in the regulation of cancer stem cells (CSCs), a subpopulation of cancer cells with self-renewal and differentiation abilities that contribute to tumor initiation, progression, and recurrence(35). HDAC1 has been shown to promote the maintenance of CSCs by regulating the expression of stemness-related genes and the self-renewal signaling pathways.

Given its critical role in cancer development and progression, HDAC1 has been targeted for cancer therapy. Several HDAC inhibitors (HDACi), including vorinostat, romidepsin, and belinostat, have been approved by the US Food and Drug Administration for the treatment of certain cancer types(36, 37). HDACi have been shown to induce cell cycle arrest, apoptosis, differentiation, and senescence in cancer cells, as well as

enhance the efficacy of other cancer therapies such as chemotherapy and radiotherapy. However, the use of HDACi in cancer therapy is still limited by their toxicity and lack of specificity for individual HDAC isoforms. HDAC1 promotes cancer progression by altering the expression of genes involved in cell growth, survival, invasion, and metastasis, as well as the regulation of CSCs. HDAC1 is a promising target for cancer therapy, and HDAC inhibitors have shown promise in preclinical models and clinical trials. Further research is needed to better understand the mechanisms of HDAC1 in cancer and to develop more specific and effective HDAC inhibitors for cancer therapy.

GATA2 is a transcription factor that plays a critical role in the development and function of hematopoietic cells, including stem cells, progenitor cells, and differentiated cells. GATA2 is a member of the GATA family of transcription factors, which bind to DNA sequences containing the consensus motif (A/T)GATA(A/G)(38). GATA2 is expressed in a tissue-specific manner, with high expression levels in hematopoietic cells and low or undetectable levels in other cell types. In hematopoiesis, GATA2 is essential for the development and maintenance of hematopoietic stem cells (HSCs) and progenitor cells(39). GATA2 regulates the expression of genes involved in cell proliferation, differentiation, survival, and migration, including cytokine receptors, transcription factors, and signaling molecules. GATA2 also interacts with other transcription factors and cofactors to form complexes that regulate gene expression. GATA2 is involved in the regulation of various stages of hematopoiesis, including self-renewal, differentiation, and lineage specification. GATA2 is required for the development of erythroid and megakaryocyte lineages and plays a crucial role in the maintenance of these cell types. GATA2 also plays a critical role in the development of lymphoid and myeloid lineages. The expression of GATA2 is regulated by various transcription factors, signaling pathways, and epigenetic mechanisms. The Notch signaling pathway is one of the critical pathways regulated by GATA2 in hematopoiesis(40). GATA2 has been shown to interact with Notch signaling components and regulate the expression of Notch target genes in HSCs and progenitor cells. The interaction between GATA2 and Notch signaling is critical for the regulation of HSC self-renewal and differentiation.

GATA2 is also involved in the regulation of various cytokine receptors and signaling pathways that are critical for hematopoiesis. GATA2 regulates the expression of cytokine receptors, such as the erythropoietin receptor (EpoR), thrombopoietin receptor (Mpl), and interleukin-7 receptor (IL-7R), and signaling molecules, such as STAT5 and Gfi1(41, 42). The regulation of these receptors and signaling pathways by GATA2 is essential for the development and function of hematopoietic cells. GATA2 interacts with various transcription factors and cofactors to regulate gene expression in hematopoiesis. GATA2 interacts with RUNX1 to regulate the expression of genes involved in megakaryopoiesis and erythropoiesis(43). GATA2 also interacts with PU.1 to regulate the differentiation of myeloid cells. The interaction between GATA2 and ETS factors has been implicated in the regulation of lymphoid cell development. The interaction between GATA2 and CBP/p300 has been shown to be critical for the regulation of hematopoietic stem and progenitor cell function(44). Mutations in GATA2 have been implicated in various hematological disorders, including leukemia, myelodysplastic syndromes (MDS), and immunodeficiency syndromes. GATA2

mutations have been identified in patients with MDS, acute myeloid leukemia (AML), and chronic myelomonocytic leukemia (CMML), and are associated with a poor prognosis(45, 46). GATA2 mutations are also found in patients with immunodeficiency syndromes, such as MonoMAC syndrome and Emberger syndrome, which are characterized by susceptibility to infections and other complications.

In addition to its role in hematopoiesis, GATA2 has been implicated in the pathogenesis of various diseases, including leukemia, myelodysplastic syndromes (MDS), and immunodeficiency syndromes. GATA2 has been implicated in the pathogenesis of several types of leukemia, including acute myeloid leukemia (AML), chronic myeloid leukemia (CML), and acute lymphoblastic leukemia (ALL)(47). In AML, GATA2 expression is often dysregulated, with some studies showing that increased expression of GATA2 is associated with poor prognosis, while others have shown the opposite. Additionally, GATA2 mutations have been identified in some cases of AML, and these mutations are thought to contribute to the development and progression of the disease. In CML, GATA2 has been shown to regulate the expression of genes involved in the progression of the disease, including BCR-ABL1(48). In ALL, GATA2 is important for the development of B-cell precursors, and abnormal GATA2 expression has been implicated in the pathogenesis of the disease. Myelodysplastic syndromes (MDS) are a group of disorders characterized by ineffective hematopoiesis and the development of abnormal blood cells. GATA2 mutations have been identified in a significant proportion of patients with MDS, and these mutations are associated with a poor prognosis. Additionally, GATA2 expression is dysregulated in some cases of MDS, with decreased expression of GATA2 associated with an increased risk of disease progression. Immunodeficiency syndromes are a group of disorders characterized by defects in the immune system, which can lead to an increased susceptibility to infections. GATA2 is important for the development and function of several immune cell types, including dendritic cells, monocytes, and B cells. Mutations in GATA2 have been identified in patients with immunodeficiency syndromes, including monocytopenia and mycobacterial infection (MonoMAC) syndrome and dendritic cell, monocyte, B, and NK lymphoid deficiency (DCML) syndrome(38, 49). These mutations are thought to impair the function of GATA2 and contribute to the development of the diseases.

Another protein that interacts with GATA2 is HDAC1, a histone deacetylase that regulates gene expression by removing acetyl groups from histones. HDAC1 has been shown to deacetylate GATA2 and repress its activity, leading to the downregulation of GATA2 target genes(50). The interaction between GATA2 and HDAC1 has been implicated in the pathogenesis of MDS and AML. In patients with MDS, GATA2 expression is downregulated, and HDAC inhibitors have been shown to restore GATA2 expression and improve hematopoiesis (13, 14). In AML, GATA2 mutations can disrupt its interaction with HDAC1 and lead to aberrant GATA2 activity.

GATA2 also interacts with other proteins involved in hematopoiesis and disease, including transcription factors such as RUNX1, PU.1, and ETS factors, as well as cofactors such as CBP and p300. GATA2 interacts with RUNX1 to regulate the expression of genes involved in megakaryopoiesis and erythropoiesis. GATA2 interacts with PU.1 to regulate the differentiation of myeloid cells(51). The

interaction between GATA2 and ETS factors has been implicated in the regulation of lymphoid cell development. The interaction between GATA2 and CBP/p300 has been shown to be critical for the regulation of hematopoietic stem and progenitor cell function.

The above literature survey shows us that SMAR1, GATA2 and HDAC1 have tremendous importance in disease biology and the diseases are related to acetylation. We hence try to visualize how does the interaction of these 3 proteins manipulate the acetylation pattern.

3. Materials and Methods:

1. Sequence download and Modelling of Proteins:

To study the atomic minutiae of interaction pattern of SMAR1-GATA2-HDAC1 complex, structure of SMAR1 and GATA2 were build and further all were docked together to get an understanding of their interaction pattern. The sequence of GATA2 and SMAR1 were downloaded from uniprot(52). A comprehensive database of protein sequences and functions called UniProt KB (Knowledge Base) offers the scientific community a convenient location to access and use data on proteins. The American Protein Information Resource (PIR), the Swiss Institute of Bioinformatics (SIB), and the European Bioinformatics Institute (EBI) have joined forces to create it. In UniProt KB, information about proteins from a variety of creatures, including humans, model organisms, and lesser-known species, is both manually annotated and automatically generated. Each protein's sequence, function, domain architecture, post-translational modifications, interaction partners, subcellular localization, and other details are all provided by the database. Data from other databases are also incorporated into UniProt KB, including information on disease-related mutations from various sources and data on protein structure from the Protein Data Bank (PDB). The public can use the database for free, and it is frequently updated to include the most recent data on proteins.

Full length SMAR1 (NP_001167010.1) and GATA2 (NP_001139133.1) structure was built with I-TASSER web-server. I-TASSER (Iterative Threading ASSEmbly Refinement) is a web server for protein structure and function prediction. It uses a hierarchical approach to predict protein structure based on threading, ab initio modeling, and iterative refinement simulations. I-TASSER generates 3D models of proteins based on their amino acid sequences and known structural templates. The server can also perform functional annotation of proteins, predicting ligand-binding sites, enzyme active sites, and protein-protein interaction interfaces(53, 54).

A. The first step in I-TASSER is threading, which involves identifying potential structural templates from the Protein Data Bank (PDB) that have sequence similarity to the target protein. This is done using the LOMETS (Local Meta-Threading-Server) algorithm, which is a consensus-based method that integrates multiple threading algorithms to improve the

accuracy of template selection(55). The LOMETS algorithm works by comparing the amino acid sequence of the target protein to the sequences of proteins in the PDB using threading algorithms. Threading is a computational method that aligns the sequence of a target protein to the sequence of a known protein structure (template) that has similar sequence and structure characteristics. Threading algorithms generate a threading score for each template, which reflects the quality of the alignment between the target and the template sequences. LOMETS uses multiple threading algorithms, including PROSPECT2, SP3, SPARKS-X, and HHsearch, to identify potential templates for the target protein. Each algorithm generates a set of templates with corresponding threading scores. LOMETS then calculates a consensus score for each template by integrating the threading scores from the different algorithms. The consensus score provides a measure of the reliability of the template and is used to rank the templates.

The advantage of using a consensus-based approach is that it improves the accuracy of template selection by reducing the influence of errors and biases in individual threading algorithms. By integrating the results from multiple threading algorithms, LOMETS can identify templates that may have been missed by any single algorithm and select the most reliable templates for further modeling. LOMETS generates a list of potential templates for the target protein, which are ranked based on their threading scores.

B. Template-based modeling: After the threading step, we constructed an initial 3D model of the target protein using a fragment-assembly method, which combines short protein fragments from the selected structural templates that are aligned to the target sequence. A library of protein fragments from the PDB database that match the amino acid sequence of the target protein were generated. The fragment length is typically between 3 to 9 residues. Next the Initial models were generated by using the selected templates to generate an initial model of the target protein by aligning the target sequence to the selected templates and building the model using homology modeling. The fragments from the library are then assembled onto the initial model using a Monte Carlo-based optimization algorithm that evaluates the compatibility of the fragment with the neighboring fragments and the overall model(56). The assembled model is further refined by optimizing the geometry, bond lengths, and angles using a molecular dynamics simulation. Finally, the model is selected based on its energy score and the consistency of the model with experimental data (if available). This generates a set of initial models, which are refined using energy minimization and molecular dynamics simulations.

C. Ab initio modeling: If no suitable template is available for the target protein, then a 3D model of the protein is generated using an ab initio modeling method. This involves generating a large number of decoys using Monte Carlo simulations and clustering the decoys based on their structural similarity. The decoy with the lowest energy score is selected as the final model.

D. Iterative refinement: After generating the initial models, I-TASSER performs iterative refinement simulations to optimize the 3D structures. This involves energy minimization and molecular dynamics simulations to improve the stereochemistry and remove steric clashes in the models. We used the CHARMM (Chemistry at HARvard Macromolecular Mechanics) force field for energy minimization to optimize the geometry, bond lengths, and angles of the model during iterative refinement(57). The CHARMM force field is a widely used molecular mechanics force field that can simulate the behavior of macromolecules such as proteins, nucleic acids, and lipids. The energy minimization algorithm uses the steepest descent and conjugate gradient methods to optimize the geometry of the model. The steepest descent method is used for the initial steps of the minimization process to quickly remove large energy barriers, while the conjugate gradient method is used for the later steps of the minimization process to refine the structure to a more stable energy minimum(58, 59). During the minimization process, the atoms in the model are moved in small steps along the direction of steepest descent or the conjugate gradient, while the forces acting on the atoms are continuously recalculated until the energy of the model reaches a minimum. The energy minimization algorithm is a computationally intensive process that requires significant computational resources. However, it is necessary for refining the structures generated by the fragment-assembly and threading algorithms used in I-TASSER.

E. Model selection: Finally, I-TASSER selects the best model based on several criteria, including energy scores, stereochemistry, and structural features. The final model is then subjected to structural validation using programs such as PROCHECK and Verify3D to ensure its quality(60).

In summary, we integrated several computational methods, including threading, fragment-assembly modeling, ab initio modeling, and iterative refinement, to generate accurate 3D models of SMAR1 and GATA2 proteins(61). The use of multiple methods and the integration of consensus-based algorithms help improve the accuracy and reliability of the predictions. The X-Ray Crystallographic structure of HDAC1 protein (PDB ID – 4BKX) was already predicted, so the coordinate file was downloaded from RCSB PDB (Protein Data Bank) site. The RCSB PDB is a public database that provides a comprehensive collection of experimentally determined 3D structures of biological macromolecules such as proteins, nucleic acids, and complex assemblies(62). It is managed and operated by the Research Collaboratory for Structural Bioinformatics (RCSB), which is a collaborative effort between Rutgers University, the University of California, San Diego, and the University of California, San Francisco.

2. *Sequence alignment solvent accessibility determination:*

To further cross check the modelling we performed sequence alignment of the top pdb hits with the respective proteins. We use ClustalW for sequence alignment. ClustalW is a widely used bioinformatics tool for multiple sequence alignment (MSA) of DNA or protein sequences(63). It was first introduced by Des Higgins in 1988 and has undergone several updates and improvements since then.

The working principle of ClustalW is based on the progressive alignment approach, which starts with aligning the two most similar sequences and then gradually adding in more sequences to the alignment(64). ClustalW uses a guide tree to cluster sequences based on their pairwise similarity scores. The guide tree is constructed by the neighbor-joining algorithm, which calculates the distances between all pairs of sequences and uses them to build a tree that reflects the evolutionary relationships between the sequences. The multiple sequence alignment is performed by iteratively aligning the sequences based on the guide tree(65). In each iteration, the most similar pairs of sequences are aligned, and the alignment is improved by adding gaps and optimizing the scoring matrix. The scoring matrix used in ClustalW is a variation of the BLOSUM matrix(66), which assigns scores to amino acid pairs based on their evolutionary frequencies.

ClustalW provides several options for fine-tuning the alignment, such as adjusting the gap opening and extension penalties, choosing the scoring matrix, and selecting the output format. It also allows users to visualize the alignment in various ways, such as a colored guide tree, a consensus sequence, and a pairwise distance matrix.

Output format was set as CLUSTAL. We used Slow method for pairwise alignment as it gives accurate result. For the pairwise alignment K-tuple(word) size was kept as 1. Window size was set as 5. Gap opening penalty and Gap extension penalty was set as 10.0 and 0.1 respectively. BLOSUM weight matrix was used. For subsequent multiple alignment Gap opening penalty and gap extension penalty was set as 10 and 0.05 respectively. Gly, Pro, Ser, Asn, Asp, Gln, Glu, Arg and Lys was selected as Hydrophilic residues. We selected BLOSUM weight matrix for sequence alignment.

To forecast the solvent accessibility of amino acid residues, we employed a two-step procedure. The SPIDER2 algorithm was used in the first stage to forecast secondary structure components and the relative solvent accessibility (RSA) of specific residues(67). The ratio of a residue's solvent accessible surface area to its maximum feasible solvent accessible surface area in a tripeptide (Ala-X-Ala) conformation is used to compute the RSA. The projected RSA values are improved in the second stage by using an upgraded SPINE-X algorithm(68). After modeling the protein's three-dimensional structure using the expected secondary structure data, SPINE-X determines RSA values based on this model. The quality of the projected models is then assessed using the improved RSA values in the final scoring function.

3. *Model quality analysis:*

Model quality was analysed using MOLPROBITY. MolProbity is a popular software tool for the validation of protein structures(69). It was developed by the Richardson Lab at Duke University and is widely used by the structural biology community to assess the quality and reliability of protein models.

MolProbity uses a range of algorithms and statistical methods to identify potential errors or inconsistencies in protein structures. It checks for geometric outliers, such as unusually high bond angles or short bond lengths, as well as for steric clashes between atoms. It also assesses the quality of the protein's electron density map and examines the positions of hydrogen atoms and water molecules(70).

The output of MolProbity includes a range of metrics and scores, such as the Ramachandran plot score, which indicates the percentage of amino acids in the protein that have acceptable backbone conformation, and the clashscore, which measures the amount of steric overlap between atoms in the protein structure.

MolProbity is widely used by researchers in the field of structural biology to validate protein structures before they are deposited in public databases such as the Protein Data Bank. It is also used to guide the process of protein structure refinement and to identify potential errors or inaccuracies that may need to be corrected.

4. *Preparation and minimization of predicted Protein structure:*

The predicted structure was further prepared and minimized using Schrodinger PRIME (Protein Refinement and Improvement using Multistate Evaluations) module(71). PRIME is a module of Schrodinger's suite of software tools that is used for protein structure refinement and improvement. It utilizes advanced computational algorithms to optimize and refine protein structures, with the goal of improving the accuracy and quality of the final model. The PRIME module works by analyzing the input protein structure and identifying potential errors or inaccuracies in the model. It then utilizes a combination of energy minimization and molecular dynamics simulations to optimize and refine the structure, while also taking into account the flexibility and dynamics of the protein. We employed a number of advanced techniques for structure refinement.

A. Side-chain optimization: PRIME uses a combination of rotamer optimization and steric clash removal to improve the placement of individual amino acid side chains within the protein structure. The optimization process involves two main steps: rotamer optimization and geometry optimization. First, the rotamer optimization step identifies the best orientation for each side chain by considering all possible rotamer conformations. This step aims to identify the most energetically favorable orientation for each side chain, given the interactions with neighboring residues and the overall protein structure. Next, the geometry optimization step adjusts the bond lengths, bond angles,

and dihedral angles of each side chain to improve the overall geometry and minimize steric clashes. This step uses molecular mechanics force fields and other algorithms to minimize the energy of the protein structure, while ensuring that the structure remains physically realistic. Side-chain optimization is important because the orientation and geometry of the side chains can have a significant impact on the overall stability and function of the protein.

B. Loop refinement: PRIME employs a loop modeling algorithm to optimize the conformation of loops within the protein structure, which are often difficult to accurately model using traditional structure determination methods. Loop refinement algorithm optimizes the conformation of these flexible regions, improving the accuracy and stability of the final protein structure. The algorithm uses a combination of molecular dynamics simulations and energy minimization to explore the conformational space of the loop and identify the lowest energy conformation. This process involves the random perturbation of the loop conformation followed by energy minimization and evaluation of the resulting structures. This iterative process continues until a stable and low-energy conformation is obtained.

   The loop refinement algorithm in PRIME is able to generate multiple conformations of the loop region, which can be further evaluated based on their energy and other criteria such as their compatibility with the experimental data. This process allows for the identification of the most stable and accurate conformation of the loop region, which can improve the overall accuracy of the protein structure.

C. Solvent optimization: PRIME uses an explicit solvent model to account for the presence of water molecules and other solvents in the protein environment, which can have a significant impact on the stability and conformation of the protein. Inaccurate modeling of solvent molecules can lead to errors in the final protein structure, and can also result in incorrect predictions of protein-ligand interactions. We used an explicit solvent model to account for the presence of water molecules and other solvents in the protein environment. This model takes into account the interactions between the protein and solvent molecules, as well as the dynamics of the solvent molecules themselves. By accurately modeling the solvent environment, PRIME can improve the accuracy and stability of the final protein structure, and can also provide more accurate predictions of protein-ligand interactions. Specifically, PRIME uses molecular dynamics simulations to optimize the positions and orientations of solvent molecules around the protein. During these simulations, the protein and solvent molecules are allowed to interact with each other, allowing the protein to adopt more stable conformations and minimizing any unfavorable interactions between the protein and solvent molecules. The resulting protein structure is therefore more accurate and better represents the true conformation of the protein in its native environment.

D. Ligand refinement: PRIME can also be used to optimize and refine protein-ligand complexes, using a combination of energy minimization and molecular dynamics simulations to improve the accuracy and stability of the complex. Overall, the PRIME module is a powerful tool for improving the accuracy and quality of protein structures and is widely used in both academic and industrial settings for protein structure refinement and optimization.

We also added the Hydrogen to the proteins and kept the pH of the proteins at 7.4.

5. *Molecular Docking:*

To analyze the stable non-covalent interaction between SMAR1, GATA2 and HDAC1, we performed docking through ZDOCK (v–3.0.2) server. ZDOCK is a Fast Fourier Transform based protein docking platform which searches all possible binding modes in the translational and rotational space amongst the two proteins and assesses each pose using an energy-based scoring function(72). ZDOCK uses a rigid-body docking approach that generates an initial set of candidate docking solutions by evaluating the shape complementarity and electrostatic interactions between the two proteins. It then uses a fast Fourier transform algorithm to refine the docking solutions and calculate their binding energy scores(73). The top-scoring solutions are further optimized and refined using Monte Carlo or molecular dynamics simulations. One of the key advantages of ZDOCK is its ability to handle large-scale protein-protein docking, making it suitable for predicting the structures of protein complexes involved in a wide range of biological processes. It has been used in numerous studies to predict the structures of protein-protein complexes, including those involved in immune recognition, virus-host interactions, and enzyme-substrate interactions. First SMAR1 and GATA2 were docked and then the complex was further docked with HDAC1.

6. *Molecular structure viewing:* The final predicted docked structure was viewed and analyzed using PyMOL. PyMOL allows us to create and manipulate high-quality 3D images of molecular structures using a variety of tools and features. It is widely used in the fields of biochemistry, molecular biology, and structural biology for a range of applications such as visualizing protein-ligand interactions, analyzing protein-protein interactions, and comparing protein structures(74). PyMOL works by importing protein structures in various file formats, such as PDB, MOL, and XYZ, and displaying them in a 3D space. Users can then manipulate the structure in a variety of ways, such as rotating, zooming, and translating. PyMOL also offers a range of rendering options to create high-quality images and videos of the molecular structures. PyMOL includes a wide range of analysis tools that allow users to calculate and display properties of the molecular structure, such as electrostatic potential, surface area, and hydrogen bonding patterns. It also has a scripting

interface that allows users to automate tasks and customize the software to their specific needs.

4. Result and Discussion:

1. *Modelling studies:*

HDAC1 structure was solved by X-RAY DIFFRACTION method having a Resolution of 3.00 Å. The R-Value Free, R-Value Work and R-Value Observed were 0.261, 0.211 and 0.213 respectively. In X-ray crystallography, the R-values (or refinement factors) are measures of the agreement between the experimental data and the model of the macromolecule being studied. There are three types of R-values commonly used in crystallography: R-value free, R-value work, and R-value observed(75).

R-value observed is the value of the R-factor calculated using all the observed diffraction data. It measures the agreement between the observed data and the model, but does not take into account the possibility of overfitting the model to the data. R-value work is the R-factor calculated using a subset of the observed data that is not used in the refinement process. Typically, about 5% of the data is used as the test set. This value gives an estimate of the accuracy of the model by assessing how well it fits the data that were not used in the refinement(76). R-value free is similar to R-value work, but uses a different set of data that was not used in the refinement process. This set is usually larger than the test set used for R-value work. The R-value free gives a more reliable estimate of the accuracy of the model, since it tests how well the model fits the data that were completely independent of the refinement process(75).

The R-values are important indicators of the quality of a crystallographic model. Low R-values indicate good agreement between the model and the data, while high R-values suggest that the model may be overfitting the data. By using different sets of data for R-value work and R-value free, we can assess the accuracy and reliability of their model, and determine whether further refinement is necessary. Since the R value of the HDAC1 is very low it points that we selected a very good structure.

The HDAC1 structure is in complex with the dimeric ELM2-SANT domain of MTA1 from the NuRD complex. We removed the ELM2-SANT domain of MTA1 using PyMOL. We also removed the ions bonded to the structure using PyMOL. In Prime module we prepped the protein structure. There we removed the steric clashes between the amino acids and flipped some of the amino acids into more acceptable rotamer form.

GATA2 and SMAR1 structures were prepared from scratch. We downloaded the sequence of GATA2 and SMAR1 from uniprot database. Table 1 highlights the sequence of both the

proteins. For both GATA2 and SMAR1 we got 4 different structures from I-TASSER. Among them we selected the best model with the highest C-score and good TM Score (Table 2). The estimated quality of the predicted model is represented by the I-TASSER C-score (Confidence score).

Table 1: Sequence of GATA2 and SMAR1

Name of Protein	Sequence (N to C terminal)
SMAR1	MMSEHDLADVVQIAVEDLSPDHPVVLENHVVTDEDEPALKRQRLEINCQDPSI KTICLRLDSIEAKLQALEATCKSLEEKLDLVTNKQHSPIQVPMVAGSPLGATQT CNKVRCVVPQTTVILNNDRQNAIVAKMEDPLSNRAPDSLENVISNAVPGRRQ NTIVVKVPGQEDSHHEDGESGSEASDSVSSCGQAGSQSIGSNVTLITLNSEEDY PNGTWLGDENNPEMRVRCAIIPSDMLHISTNCRTAEKMALTLLDYLFHREVQA VSNLSGQGKHGKKQLDPLTIYGIRCHLFYKFGITESDWYRIKQSIDSKCRTAWR RKQRGQSLAVKSFSRRTPNSSSYCPSEPMMSTPPPASELPQPQPQPQPQALHYALA NAQQVQIHQIGEDGQVQVGHLHIAQVPQGEQVQITQDSEGNLQIHHVGQDGQ LLEATRIPCLLAPSVFKASSGQVLQGAQLIAVASSDPAAAGVDGSPLQGSDIQV QYVQLAPVSDHTAGAQTAEALQPTLQPEMQLEHGAIQIQ
GATA2	MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNH LDSQGNPYYANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGK AALSAAAAHHHNPWTVSPFSKTPLHPSAAGGPGGPLSVYPGAGGGSGGGSGS SVASLTPTAAHSGSHLFGFPPTPPKEVSPDPSTTGAASPASSSAGGSAARGEDK DGVKYQVSLTESMKMESGSPLRPGLATMGTQPATHHPIPTYPSYVPAAAHDYS SGLFHPGGFLGGPASSFTPKQRSKARSCSEGRECVNCGATATPLWRRDGTGHY LCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTCCANCQTTTTTLWRRNAN GDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSNKSKKSKKGAECFEE LSKCMQEKSSPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSSLSFGHPH PSSMVTAMG

Table 2: Model attributes of SMAR1 and GATA2

Name	C-score	Exp.TM-Score	Exp.RMSD	No.of decoys	Cluster density
SMAR1	1.70	0.51±0.15	11.0±4.6	600	0.0791
GATA2	-0.49	0.65±0.13	8.3±4.5	600	0.2500

Higher scores reflect greater confidence in the model; the normalized value ranges from -5 to 2. The importance of threading template alignments and the convergence parameters of the structure assembly simulations are the two main determinants of the C-score. A structural similarity metric used to compare two protein structures is called the TM-score (Template Modeling score). It has a scale of 0 to 1, with 1 denoting complete resemblance between two structures. By comparing the predicted models in I-TASSER to recognized structures in the PDB database, the TM-score is utilized to evaluate the precision of the predicted models.

Next to verify the model quality we submitted the structures to Molprobity. Molprobity analysed the bond length and angle geometry analysis by mp_geo algorithm. Performed Ramachandran analysis and make plots by ramalyze algorithm. It also performed rotamer (rotalyze), Cβ deviation, cis-peptide (omegalyze), and CαBLAM analysis. The model quality analysis report is attached in Table 3, Table 4, appendix 1 and Appendix 2. We could see that for GATA2 there were 82 outliers in Ramachandran Plot.

Table 3: Model Quality analysis of GATA2 predicted by I-Tasser

Protein Geometry	Poor rotamers	45	11.94%	Goal: <0.3%
	Favored rotamers	276	73.21%	Goal: >98%
	Ramachandran outliers	82	17.15%	Goal: <0.05%
	Ramachandran favored	263	55.02%	Goal: >98%
	Rama distribution Z-score	-6.91 ± 0.28		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	35	8.14%	Goal: 0
	Bad bonds:	3 / 3663	0.08%	Goal: 0%
	Bad angles:	105 / 4989	2.10%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	3 / 54	5.56%	Expected: ≤1 per chain, or ≤5%
	Twisted Peptides:	36 / 479	7.52%	Goal: 0
Low-resolution Criteria	CaBLAM outliers	133	27.9%	Goal: <1.0%
	CA Geometry outliers	40	8.40%	Goal: <0.5%
Additional validations	Chiral handedness swaps	2/463	0.43%	See Chiral volume report for details
	Tetrahedral geometry outliers	1		

Table 4: Model Quality analysis of GATA2 after refinement

Protein Geometry	Poor rotamers	29	7.69%	Goal: <0.3%
	Favored rotamers	312	82.76%	Goal: >98%
	Ramachandran outliers	42	8.79%	Goal: <0.05%
	Ramachandran favored	320	66.95%	Goal: >98%
	Rama distribution Z-score	-5.70 ± 0.30		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	34	7.91%	Goal: 0
	Bad bonds:	0 / 3663	0.00%	Goal: 0%
	Bad angles:	107 / 4989	2.14%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	4 / 54	7.41%	Expected: ≤1 per chain, or ≤5%
	Twisted Peptides:	22 / 479	4.59%	Goal: 0
Low-resolution Criteria	CaBLAM outliers	75	15.8%	Goal: <1.0%
	CA Geometry outliers	39	8.19%	Goal: <0.5%
Additional validations	Tetrahedral geometry outliers	3		

Next, we prepared the protein in protein preparation wizard of schrodinger and optimised the protein in PRIME module. Now again we analysed the structure. We could see that the structure becomes much better. Compared to the modelled structure every parameter now shows better values. The Ramachandran outliers dropped from 82 to 42. The Ramachandran polt (Figure 1-2) shows much better convergence of the amino acids. To further check if the protein has really folded and the Ramachandran plot is optimum we analysed the Ramachandran Z score of the protein. The Ramachandran Z score < 2 suggests good modelling of the protein. In case of GATA2 we found the Z score to be -5.7 which indicates very clean structure (Appendix 3).

Next, we prepared the SMAR1 protein in protein preparation wizard of schrodinger and optimised the protein in PRIME module. We could see that the structure becomes much better. Compared to the modelled SMAR1 structure every parameter now shows better values (table 5, Table 6 and Appendix 4 and Appendix 5). The Poor rotamer dropped from 68 to 52 and the Ramachandran outliers dropped from 82 to 42. The protein geometry after refimenet has much better optimised value that points to the good structure refinement. There is still scope of optimization but that is beyond the scope pof this thesis. For this we have to simulate the protein structures for atleast 10 ns to remove all the bad geometry. The

Ramachandran polt (Figure 3-4) shows much better convergence of the amino acids. To further check if the protein has really folded and the Ramachandran plot is optimum, we analysed the Ramachandran Z score of the protein. The Ramachandran Z score -5.61 suggests a very good modelling of the protein (Appendix 6).

To sum up we downloaded the structure of HDAC1 and modelled the structure of SMAR1 and GATA2. Then we further refined the structure to relieve the bad bonds, poor rotamer and fold the protein to an optimum state.

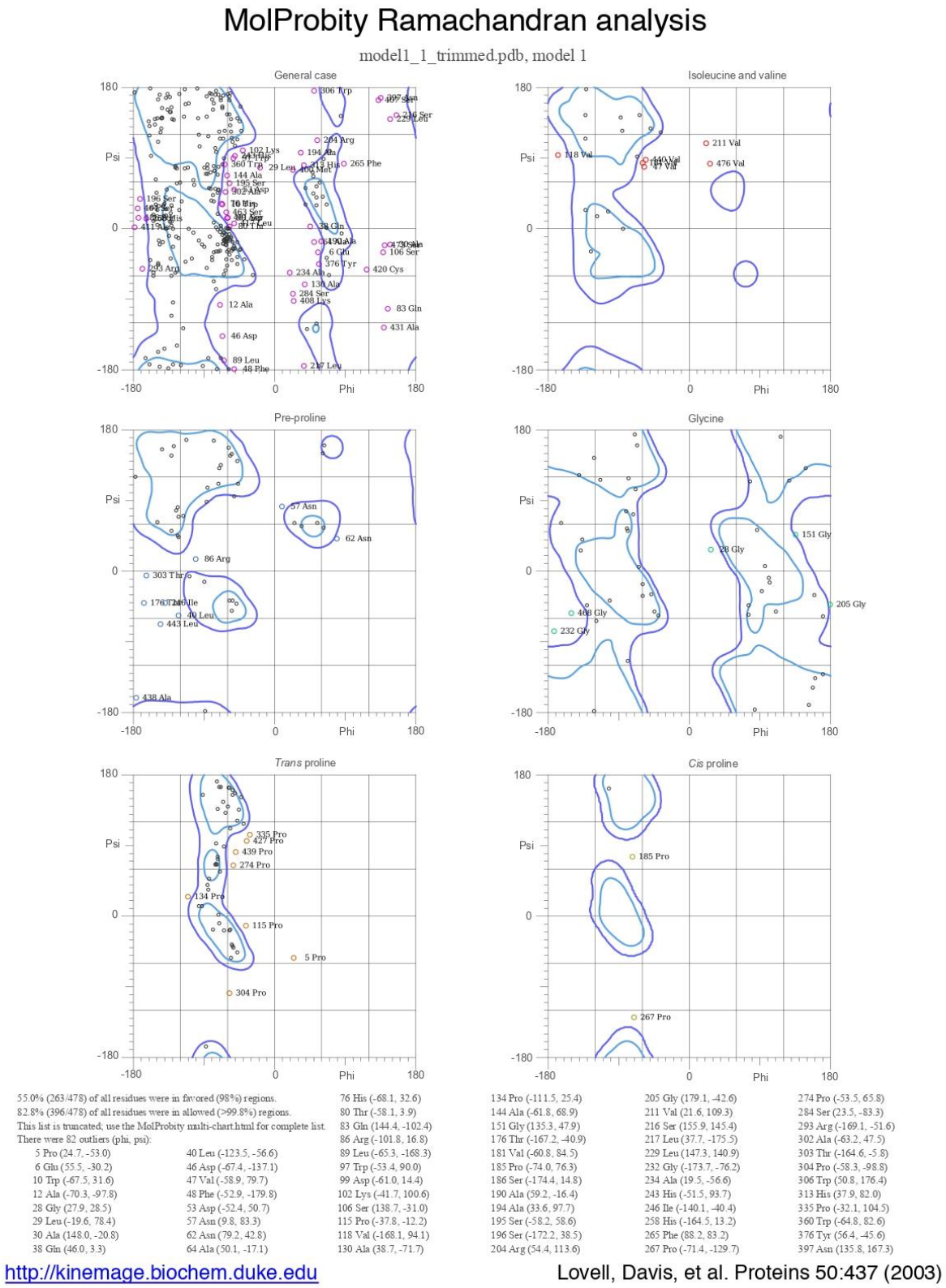


Figure 2. Ramachandran plot of GATA2 after refinement

Table 5: Model Quality analysis of SMAR1 predicted by I-Tasser

Protein Geometry	Poor rotamers	90	19.03%	Goal: <0.3%
	Favored rotamers	312	65.96%	Goal: >98%
	Ramachandran outliers	68	12.45%	Goal: <0.05%
	Ramachandran favored	349	63.92%	Goal: >98%
	Rama distribution Z-score	-6.62 ± 0.26		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	43	8.38%	Goal: 0
	Bad bonds:	1 / 4252	0.02%	Goal: 0%
	Bad angles:	146 / 5788	2.52%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 34	0.00%	Expected: ≤1 per chain, or ≤5%
	Cis nonProlines:	3 / 513	0.58%	Goal: <0.05%
	Twisted Peptides:	56 / 547	10.24%	Goal: 0
Low-resolution Criteria	CaBLAM outliers	100	18.4%	Goal: <1.0%
	CA Geometry outliers	41	7.54%	Goal: <0.5%
Additional validations	Tetrahedral geometry outliers	3		

Table 6: Model Quality analysis of GATA2 after refinement

Protein Geometry	Poor rotamers	33	7.57%	Goal: <0.3%
	Favored rotamers	351	80.50%	Goal: >98%
	Ramachandran outliers	52	10.70%	Goal: <0.05%
	Ramachandran favored	351	72.22%	Goal: >98%
	Rama distribution Z-score	-5.67 ± 0.29		Goal: abs(Z score) < 2
	Cβ deviations >0.25Å	52	10.66%	Goal: 0
	Bad bonds:	0 / 3917	0.00%	Goal: 0%
	Bad angles:	100 / 5313	1.88%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	0 / 33	0.00%	Expected: ≤1 per chain, or ≤5%

	Cis nonProlines:	1 / 472	0.21%	Goal: <0.05%
	Twisted Peptides:	10 / 505	1.98%	Goal: 0
Low-resolution Criteria	CaBLAM outliers	87	16.9%	Goal: <1.0%
	CA Geometry outliers	35	6.80%	Goal: <0.5%
Additional validations	Tetrahedral geometry outliers	3		

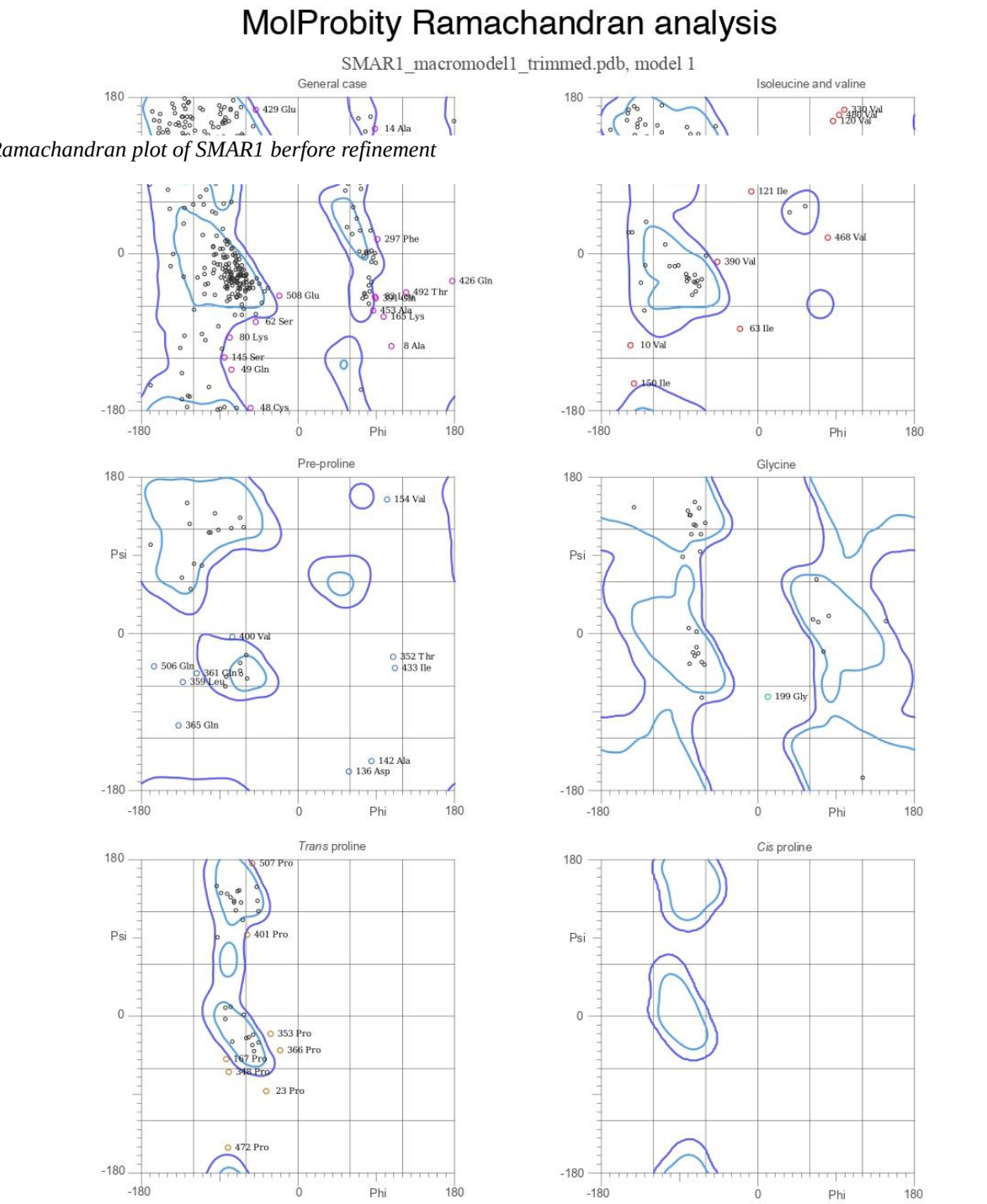


Figure 3. Ramachandran plot of SMAR1 berfore refinement

Figure 4. Ramachandran plot of SMAR1 after refinement

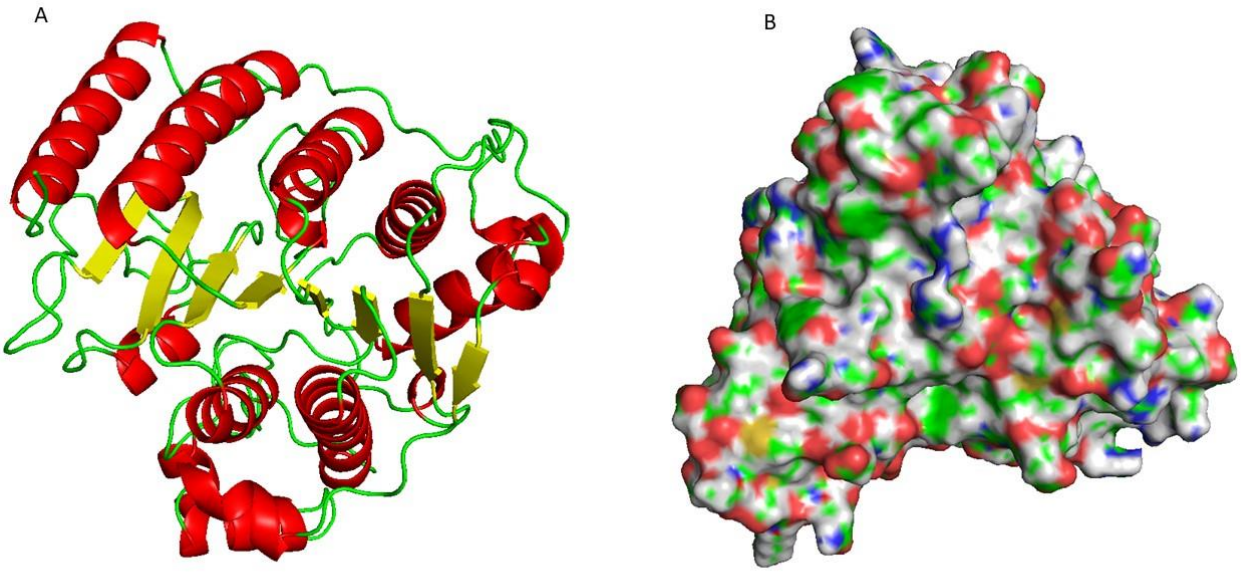


Figure 5. Structure of HDAC1. (A). Cartoon structure highlighting the secondary strucures. (B). THe surface view of HDAC1

The HDAC1 protein is a globular protein with 15 helices and 8 β strands. The β sheets are much smaller in length. Surprisingly the sheets are all lined up side by side giving a compact structure (Figure 5).

The GATA2 structure that we modelled on the other hand contained very little folded conformation. The maximum of the area is in loop. This is because the GATA2 structure contains a significant number of loops. This is because GATA2 is a transcription factor protein that binds to DNA and regulates gene expression. The DNA-binding domain of GATA2 is typically composed of short alpha-helices and beta-strands that are connected by loops. These loops are essential for the protein to achieve the correct conformation and to interact with the DNA in a specific manner (figure 6).

In addition to the DNA-binding domain, GATA2 also contains other functional domains, such as a zinc finger domain and a transactivation domain. These domains also contain loops that are important for their respective functions. For example, the zinc finger domain contains loops that coordinate a zinc ion, which is critical for its DNA-binding activity.

Overall, the loop regions in GATA2 are crucial for the protein's function and stability. They help to maintain the correct three-dimensional structure of the

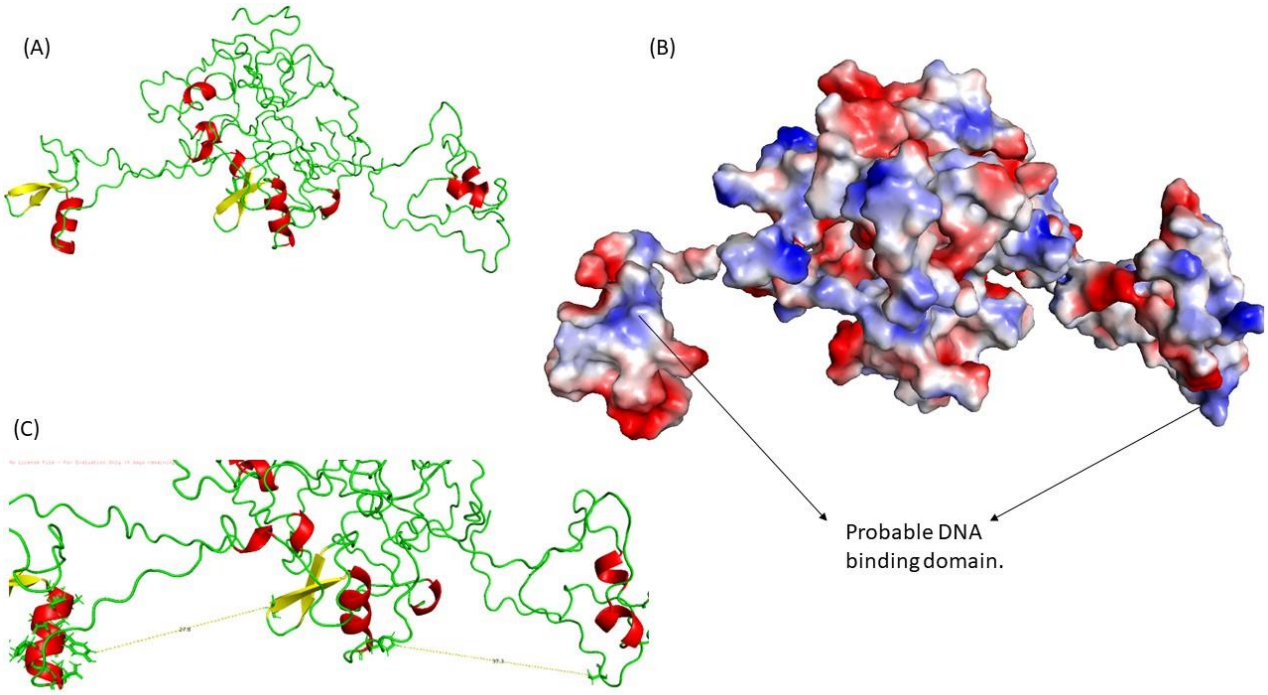


Figure 6. The structure of GATA2. (A) the structureis majorly covered with loops. (B). electrostatic of the GATA2. The claw like two appendages might be DNA binding domain. As it can hold DNA and is Positively charged which is essential to bind to phosphate backbone of DNA. (C) the diameter of the two claw is 28 Å and 37 Å, which is sufficient to accommodate DNA.

protein and allow it to interact with other molecules, such as DNA, proteins, and cofactors. As we can see in Figure 6.C the diameter of the two claw is ~28 Å and ~37 Å, which is sufficient to accommodate double stranded B-DNA. The loop also ensures that it may tight the grip by moving the either arm accordingly.

The SMAR1 was again found to assumes a globular structure. SMAR1 is a 678 amino acid protein that belongs to the ARID (AT-rich interaction domain) family of DNA-binding proteins. It contains an N-terminal ARID domain (amino acids 67-180) that recognizes and binds to AT-rich regions of DNA, a central MAR (matrix attachment region) binding domain (amino acids 238-419), and a C-terminal proline-rich domain (amino acids 586-633) that interacts with other proteins.

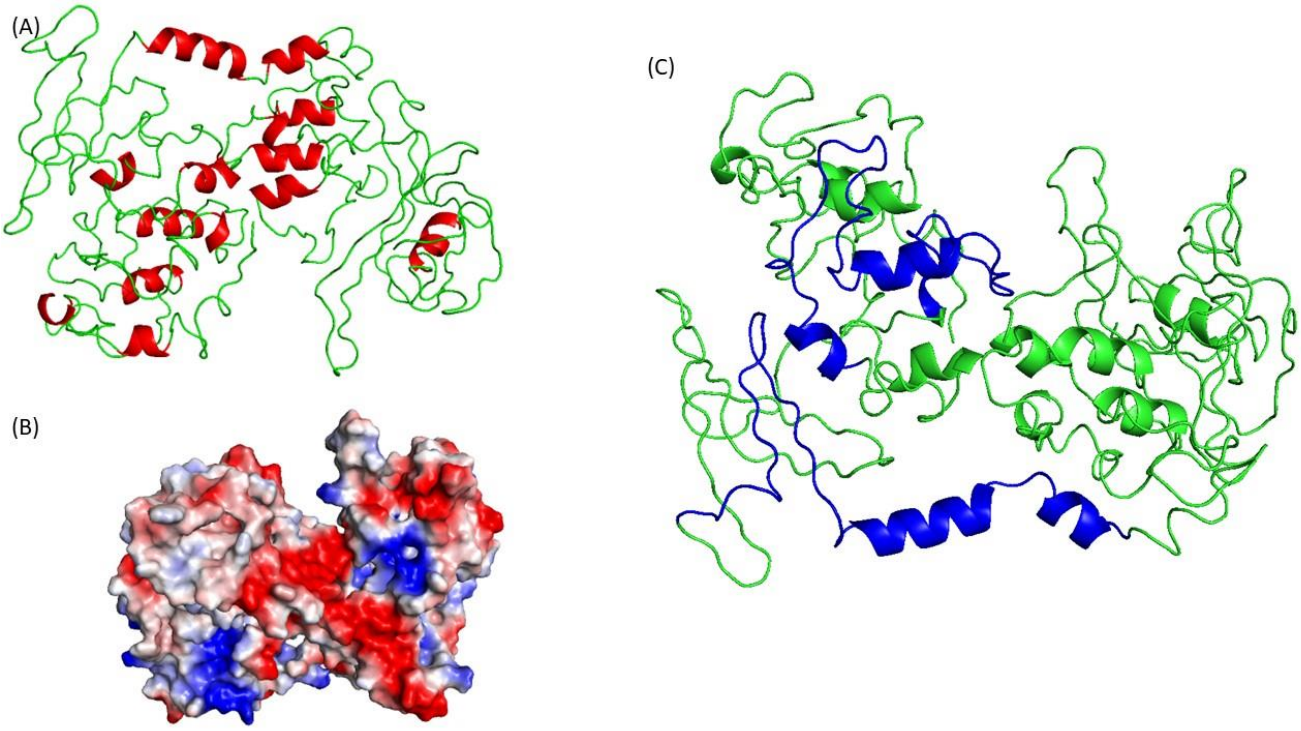


Figure 7. The structure of SMAR1. (A) Just like any SAND domain protein the structure is devoid of any β sheets. (B). Vacuum Electrostatic of the SMAR1.. (C) the blue coloured domain is the conserved MAR domain, which has DNA binding ability. This domain is the most conserved sequence in SMAR1.

The MAR binding domain of SMAR1 contains two highly conserved regions (amino acids 238-284 and 364-419) (Figure 7) that are involved in DNA binding and protein-protein interactions. We can see that the conserved domain is partially structured as it contains the helices. The region between the two conserved motifs is less conserved and contains a predicted helix-turn-helix motif. SMAR1 has also been shown to form oligomers through its MAR-binding domain, which may contribute to its function in chromatin organization and gene regulation.

In addition to the conserved domains, SMAR1 contains several post-translational modification sites, including serine, threonine, and tyrosine residues that can be phosphorylated, lysine residues that can be acetylated, and arginine residues that can be methylated. These modifications may regulate the activity and stability of SMAR1, as well as its interactions with other proteins and DNA.

SMAR1 belongs to the SAND family of proteins. It is characterized by the presence of a highly conserved SAND domain, which is evident as like any SAND domain containing protein SMAR1 contains several helices but no beta sheets (Figure 7.A). The SAND domain of SMAR1 is responsible for DNA binding and is essential for its regulatory functions in the cell.

It is seen that though both GATA2 and SMAR1 are DNA bonding proteins there are some remarkable difference between these two proteins which is evident from their structure. GATA2 contains both alpha helices and beta sheets, which form a compact structure that facilitates its DNA-binding activity. In contrast, SMAR1 contains predominantly alpha helices and lacks beta sheets, which suggests that it may adopt a more extended or flexible

structure. GATA2 has been shown to be a stable protein that can tolerate mutations in its zinc finger domains without significant loss of function. In contrast, SMAR1 is highly sensitive to point mutations, which can destabilize its structure and impair its binding to chromatin which is evident from its lack of secondary strcutres and presence of long loops.

2. *Sequence alignment and solvent accessibility determination:*

Next, we employed sequence alignment to verufy if the correct sequnces have been identified for GATA2 modelling. For this employed CLUSTALW bioinformatics tool. The sequence of Proteins selected for homology modelling are HEP200 protein of *Cylindrotheca fusiformis* (PDB ID: 2NBI, A chain), SUPERKILLER PROTEIN 3 of *S. cerevisiae* (PDB ID: 4BUJ, B chain), FATTY ACID SYNTHASE SUBUNIT BETA of *SACCHAROMYCES CEREVISIAE* (PDB ID: 2VKZ, G chain), (PDB ID: 2VKZ, G chain), Acetyl-CoA carboxylase 1 of *Homo sapiens* (PDB ID: 6G2D, C chain), Inositol 1,4,5-trisphosphate receptor type 1 of *Rattus norvegicus* (PDB ID: 3JAV, A chain), Serine/threonine-protein kinase ATR of Homo sapien (PDB ID: 5YZ0, A chain), Clathrin heavy chain of Bos taurus (PDB ID: 1XI4, A chain), Pre-mRNA-processing-splicing factor 8 of Homo sapiens (PDB ID: 5XJC, A chain), Serine/threonine-protein kinase MEC1|Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (PDB ID: 5X6O, C chain), FATTY ACID SYNTHASE of SUS SCROFA (PDB ID: 2VJ8, B chain). The alignment report is shown next in Table 7.

Table 7: Sequence alignment with top 10 pdb hits of GATA2.

```
GATA2      ------------------------------------------------------------
2NBI_1|Cha ------------------------------------------------------------
4BUJ_2|Cha --------------------GPDSMSDIKQLLKEAKQELTN------------RDYEETI
5YZ0_1|Cha MGE-----------------HGLELASMIPALRELGSATP-------------EEYNTVV
5X6O_1|Cha MES-----------------HVKYLDELILAIKDLNSGV-----------------DSKV
1XI4_1|Cha MAQILP----------------------------------------------IRFQEHL
3JAV_1|Cha ------------------------MSDKMSSFLHIGDICS--------------LYAEGS
2VZ8_1|Cha MEE----------------VVIAGMSGKLPESENLEEFWANLIGGVDMVTADDRRWKAGL
6G2D_1|Cha MAHHHHHHH------HHHGSTSGSGEQKLISEEDLGSTSGS---------GDYKDDDDKL
2VKZ_2|Cha MDA--------------YSTRPLTLSH---------------------------------
5XJC_1|Cha MAGVFPYRGPGNPVPGPLAPLPDYMSE-----EKLQEKARK------WQQLQAKRYAEKR
```

We can see the alignment does not bear tremendous similarity and as a result non homologous part were calculated by ab initio method. We also determined the solvent accessibility of the individual amino acid. The degree to which an amino acid residue is exposed to the solvent environment rather than being buried deep inside the protein is referred to as the amino acid's solvent accessibility. Because it offers details about the residue's immediate surroundings, which can be used to forecast the residue's function and interactions with other molecules, it is a crucial parameter in protein modeling.

Generally speaking, amino acids with greater solvent accessibility are found closer to the surface of proteins and are more likely to interact with other proteins or small molecules. Involvement in protein-protein interactions, binding to ligands, and immune system recognition are a few examples of these interactions. On the other hand, amino acids that are more deeply embedded in the protein structure are not prone to interact with other proteins. Cross checking of the amino acid with solved structure of GATA2 (Table 8) signalled that the protein folded properly and the amino acids showing higher solvent accessibility were placed in the surface of the proteins.

Table 8: Solvent accessibility of the GATA2.

Sequence	MEVAPEQPRWMAHPAVLNAQHPDSHHPGLAHNYMEPAQLLPPDEVDVFFNHLDSQGNP YYANPAHARARVSYSPAHARLTGGQMCRPHLLHSPGLPWLDGGKAALSAAAAHHHNPW TVSPFSKTPLHPSAAGGPGGPLSVYPGAGGGSGGGSGSSVASLTPTAAHSGSHLFGFP PTPPKEVSPDPSTTGAASPASSSAGGSAARGEDKDGVKYQVSLTESMKMESGSPLRPG LATMGTQPATHHPIPTYPSYVPAAAHDYSSGLFHPGGFLGGPASSFTPKQRSKARSCS EGRECVNCGATATPLWRRDGTGHYLCNACGLYHKMNGQNRPLIKPKRRLSAARRAGTC CANCQTTTTTLWRRNANGDPVCNACGLYYKLHNVNRPLTMKKEGIQTRNRKMSNKSKK SKKGAECFEELSKCMQEKSSPFSAAALAGHMAPVGHLPPFSHSGHILPTPTPIHPSSS LSFGHPHPSSMVTAMG
Prediction	664455433123233223343344433433332333231133730320022234433 2 322323333343323333443443222313231332101234343333333211211 2 333121321212333333333231311222233333333333333322323123323 233244332333333331222222222213333333242112223313322123223 2 212232212322223221232133313232223223221223123233444344443 4 433323314233132133244243314433134334433443334333322443333 2 033132341211133563320000000012234341134034741442424455655 4 555445455345434444443444323343334333334334434344332323434 4334123112144147

0 (buried residue) to **9** (highly exposed residue)

Similar sequence alignment (Data not shown) and solvent accessibility of SMAR1 corroborated with its solved structure (table 9). This proves that our modelled protein has folded properly.

Table 9: **Solvent accessibility of the SMAR1.**

Sequence	MMSEHDLADVVQIAVEDLSPDHPVVLENHVVTDEDEPALKRQRLEINCQDPSIKTICL RLDSIEAKLQALEATCKSLEEKLDLVTNKQHSPIQVPMVAGSPLGATQTCNKVRCVVP QTTVILNNDRQNAIVAKMEDPLSNRAPDSLENVISNAVPGRRQNTIVVKVPGQEDSHH EDGESGSEASDSVSSCGQAGSQSIGSNVTLITLNSEEDYPNGTWLGDENNPEMRVRCA IIPSDMLHISTNCRTAEKMALTLLDYLFHREVQAVSNLSGQGKHGKKQLDPLTIYGIR CHLFYKFGITESDWYRIKQSIDSKCRTAWRRKQRGQSLAVKSFSRRTPNSSSYCPSEP MMSTPPPASELPQPQPQPQALHYALANAQQVQIHQIGEDGQVQVGHLHIAQVPQGEQV QITQDSEGNLQIHHVGQDGQLLEATRIPCLLAPSVFKASSGQVLQGAQLIAVASSDPA AAGVDGSPLQGSDIQVQYVQLAPVSDHTAGAQTAEALQPTLQPEMQLEHGAIQIQ
Prediction	735543013003000330347333314443345664343434424251543303300 2 203310121043043304203330430341042034334333323324443333353 2 443313223110012244444434434444345344543443235414443342010 20 134633445746444433433443343444422330100003145313402100344 2 231203010334101201320330330000001101234130133122424334431 3 101000000000231304342033034303440340133343334022331344334 3 443334433433323344245333310220123233031332345232231343323 0 241363430301434523020210244244242243424333433342334313433 3 433434343321311310201000232223321324315255031211310312134 3 334324303420334130542315288749852168974523458974569235 7

0 (buried residue) to **9** (highly exposed residue)

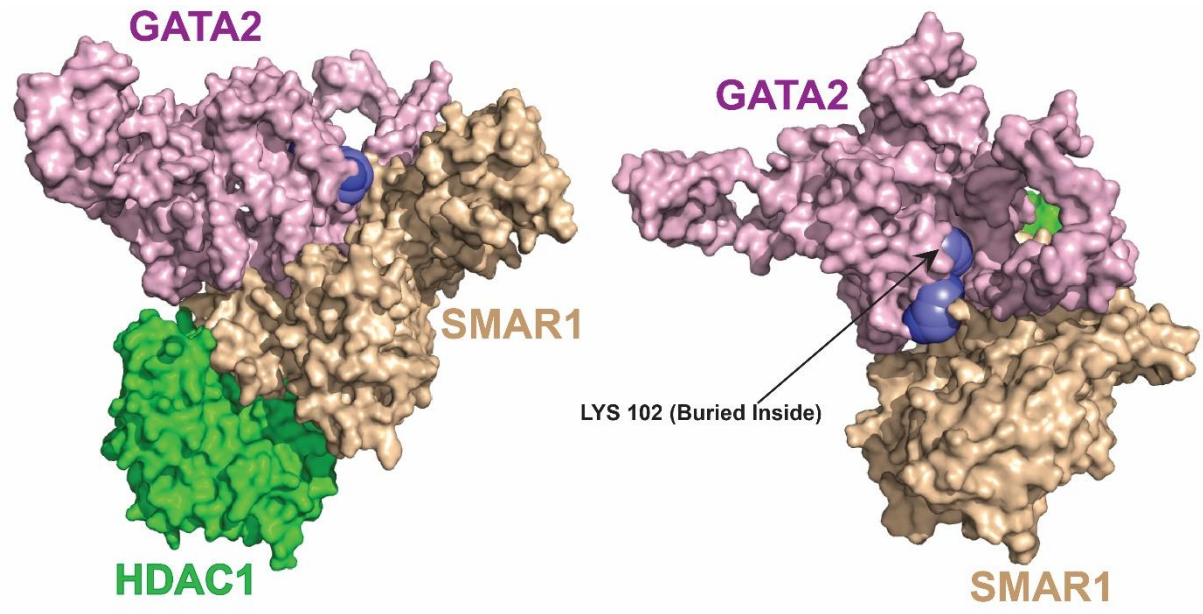


Figure 9. Upon binding with HDAC1 the Lys 102 of GATA2 of gets buried inside.

3. *Molecular docking:* To analyse the binding of these three proteins and to find atomic minitue, we docked them together. First, we docked SMAR1 with GATA2. We found that that

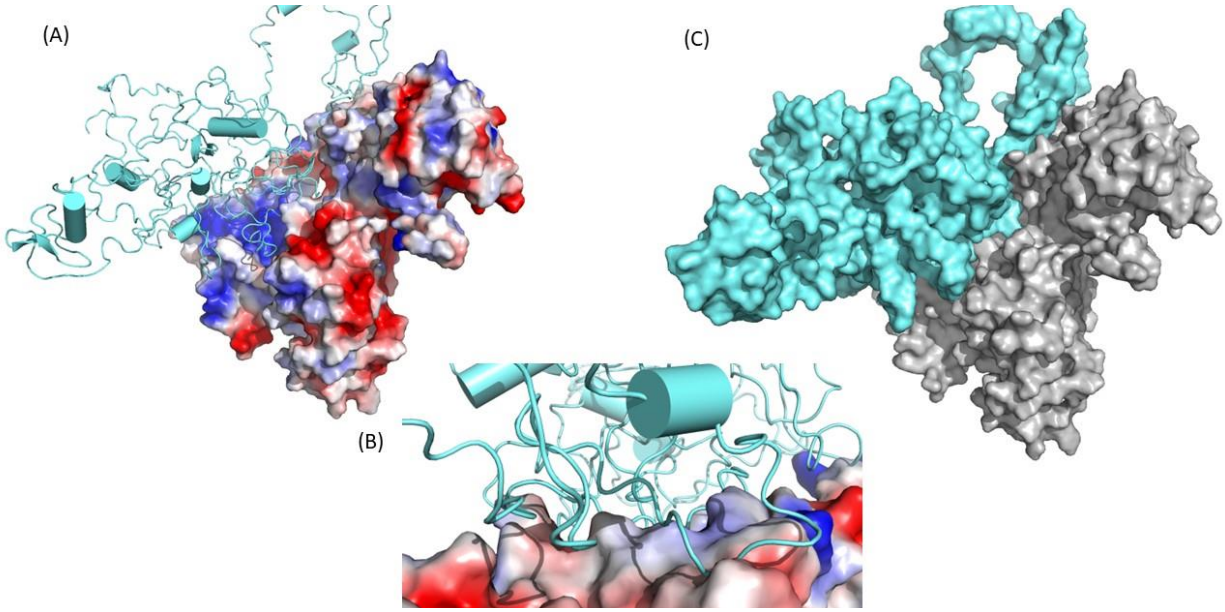


Figure 8. Docked model of SMAR1 (Cyan) and GATA2 (Grey). (A)-(B). The SMAR1 (cyan) bonded to GATA2 (surface). (C). Surface model structure of SMAR1 and GATA2.

residue 192-351 of SMAR1 is binding to GATA2 (Figure 8). The SMAR1 was binding to GATA2 using protein binding grooves as seen in the figure 7. The N-terminal Zinc Finger domain of GATA2 contains many Lysine at position 102, 123, 179, 208, 212, 222, 281 and 285. Among them the major acetylation site is Lys 102, 123, 281 & 285. From the docking it was observed that on interacting with SMAR1, the Lys 102, 123 and 179 becomes unavailable for acetylation. Lys 102 gets buried inside (Figure 9), Lys 123 forms a salt bridge with Gln 192 of SMAR1 and Lys 179 positions itself inside a cleft interacting with SMAR1 (Figure 10).

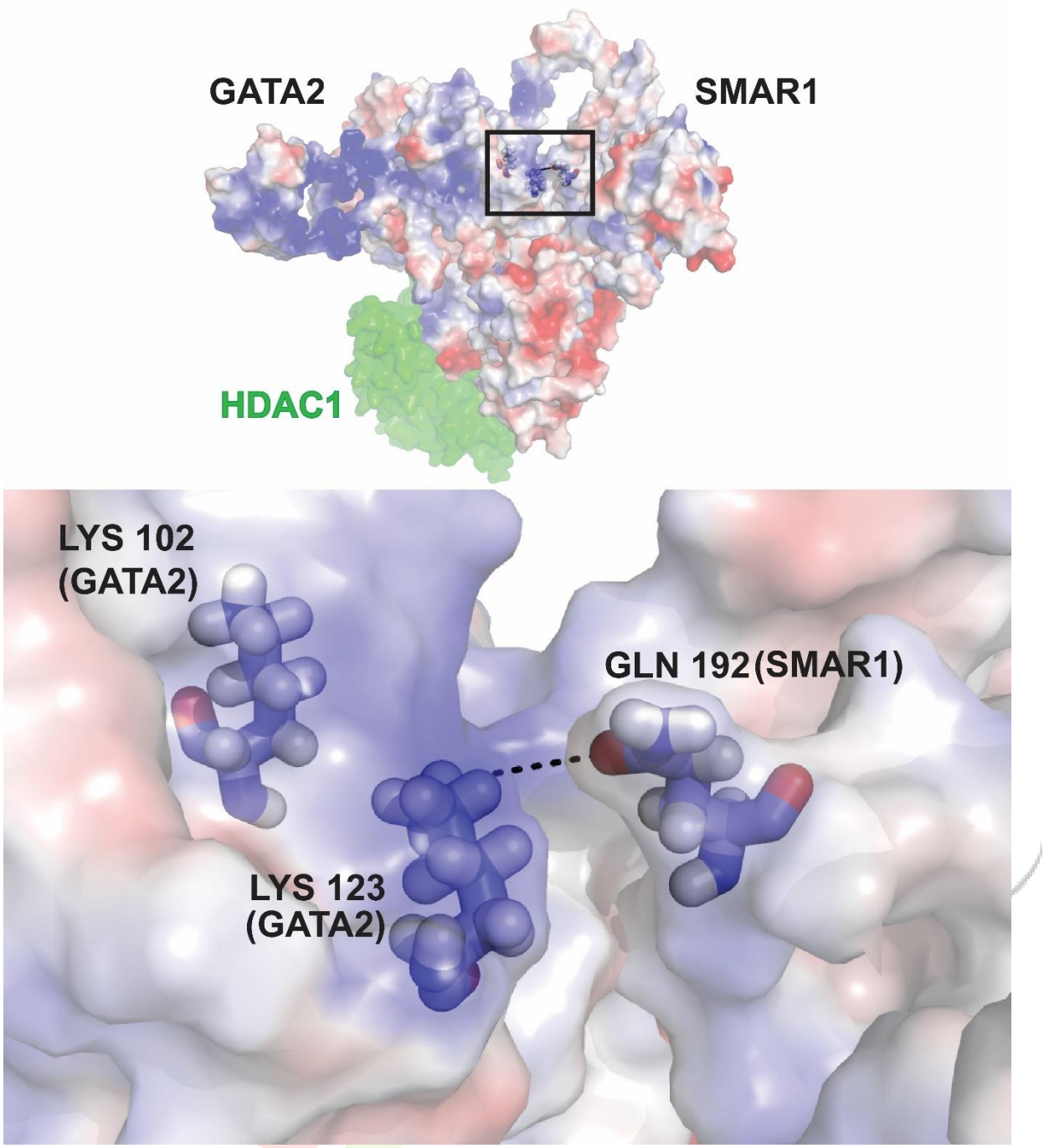
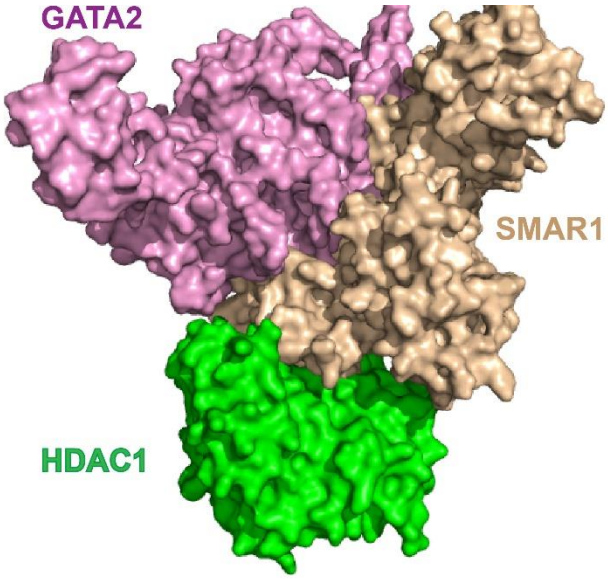


Figure 10. Docking of HDAC1 with SMAR1-GATA2 complex, shows Lys 123 of GATA2 is in close proximity to Gln192 of SMAR1.

Next, we docked the SMAR1-GATA2 complex with HDAC1. HDAC1 used 271-306 residues of SMAR1 to bind. It did not bond to GATA2 (Figure 12). SMAR1 acted as the docking site for both the proteins. Figure 12 shows all the Lysine of GATA2. Further on docking HDAC1 with SMAR1-GATA2 complex, it is seen that Lys 222 of GATA2 is in close proximity to Gln 26 of HDAC1.

Figure 11. SMAR1-GATA2-HDAC1 complex. HDAC1 and GATA2 docked onto SMAR1.



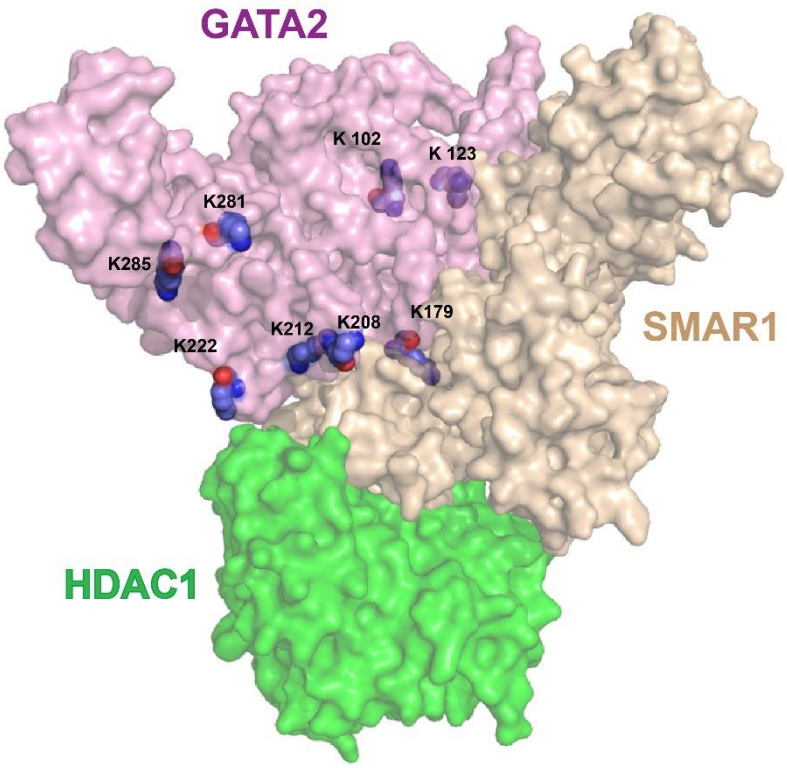


Figure 12. All the Lysine residues of GATA2.

That leaves only Lys 281 and Lys 285 on surface available for acetylation. From the result it is clear that on interacting with SMAR1 and HDAC1, not enough Lysine of GATA2 is present to be acetylated. Thus, sufficient acetylation fails to occur and GATA2 in presence of SMAR1 and HDAC1 gets very weakly acetylated.

Calnexin is an ER resident protein with calcium binding ability. It has known functions in glycoprotein folding and maturation. Cumulative evidences indicate the implication of calnexin in apoptosis induced by ER stress. Calnexin gene silencing in lung cancer cell line was shown to decrease cancer cell survival leading to effective chemotherapy. MAR binding protein SMAR1, established to have both tumor suppressor as well as immuno-modulatory

functions. We speculated that apart from its tumor suppressor function, SMAR1 might also be involved in immunosurveillance of cancer cells.

Earlier results depicted that SMAR1 increases the enrichment of both GATA2 and HDAC1 at calnexin promoter that they might interact with each other and form a repressor complex. However, no reports are available showing its interaction with HDAC1. SMAR1 is known to interact with HDAC1, but its interaction with GATA2 is unknown. We hypothesize that SMAR1 might form a triple complex with GATA2 and HDAC1 resulting in deacetylation of GATA2. We then checked the interaction between SMAR1, GATA2 and HDAC1. GATA2 Acts as an Activator of Calnexin in the Absence of SMAR1. GATA2 is known to act as an activator under acetylated condition, this acetylation is generally carried out by p300, an important member of HAT family of proteins. We try to establish that SMAR1 forms a triple complex with GATA2 and HDAC1. In the presence of SMAR1, there is reduction in acetylation of GATA2. So, we further want to check how SMAR1 and HDAC1 helps in the weak acetylation of GATA2.

5. CONCLUSION:

Generally, SMAR1 binds to calnexin as an activator protein but here we deduce that the three-protein complex might negatively regulates its activity by modulating the acetylation sites which is very important for gene regulation. This three-molecule complex might change the way calnexin expression is regulated which would be important for anticancer drug design.

Through this project we learned why is it important to know about Calnexin, GATA2, SMAR1 and HDAC1. We performed the literature study to analyse their role in modulating and prognosis of various diseases. We gathered information about various database and found how to hover these databases. We modelled the protein and through in-silico docking, it was discovered that SMAR1's residues 192–351 interact to GATA2. Numerous Lysines may be found in GATA2's N-terminal Zinc Finger domain at positions 102, 123, 179, 208, 212, 222, 281 and 285. The main acetylation sites among them are Lys 102, 123, 281 & 28. From the docking, it was seen that the Lys 102, 123, and 179 become inaccessible for acetylation when they interact with SMAR1. Lys 102 becomes encased, Lys 123 and SMAR1's Gln 192 create a salt bridge, and Lys 179 places itself inside a cleft to interact with SMAR1. Further docking of HDAC1 with the SMAR1-GATA2 complex reveals that GATA2's Lys 222 and HDAC1's Gln 26 are in close proximity. Only Lys 281 and Lys 285 are now available for acetylation on the surface. The outcome makes it obvious that not enough GATA2 lysine is present to be acetylated when it interacts with SMAR1 and HDAC1. Due of this insufficient acetylation, GATA2 is only very faintly acetylated when SMAR1 and HDAC1 are present.

This discovery provided information about the context-dependent transformation of an activator like GATA2 into a repressor.

C. References:

1. Filipeanu CM. Chapter Eleven - Temperature-Sensitive Intracellular Traffic of α2C-Adrenergic Receptor. In: Wu G, editor. Progress in Molecular Biology and Translational Science. 132: Academic Press; 2015. p. 245-65.
2. Rosenbaum EE, Hardie RC, Colley NJ. Calnexin is essential for rhodopsin maturation, Ca2+ regulation, and photoreceptor cell survival. Neuron. 2006;49(2):229-41.
3. Michalak M, Robert Parker JM, Opas M. Ca2+ signaling and calcium binding chaperones of the endoplasmic reticulum. Cell calcium. 2002;32(5-6):269-78.
4. Venkatesan A, Satin LS, Raghavan M. Roles of Calreticulin in Protein Folding, Immunity, Calcium Signaling and Cell Transformation. Progress in molecular and subcellular biology. 2021;59:145-62.
5. Raposo CD, Canelas AB, Barros MT. Human Lectins, Their Carbohydrate Affinities and Where to Find Them. Biomolecules. 2021;11(2).
6. Ryan D, Carberry S, Murphy Á C, Lindner AU, Fay J, Hector S, et al. Calnexin, an ER stress-induced protein, is a prognostic marker and potential therapeutic target in colorectal cancer. Journal of translational medicine. 2016;14(1):196.
7. Delom F, Emadali A, Cocolakis E, Lebrun JJ, Nantel A, Chevet E. Calnexin-dependent regulation of tunicamycin-induced apoptosis in breast carcinoma MCF-7 cells. Cell Death & Differentiation. 2007;14(3):586-96.
8. Kuang XL, Liu F, Chen H, Li Y, Liu Y, Xiao J, et al. Reductions of the components of the calreticulin/calnexin quality-control system by proteasome inhibitors and their relevance in a rodent model of Parkinson's disease. Journal of neuroscience research. 2014;92(10):1319-29.
9. Dubuisson J, Rice CM. Hepatitis C virus glycoprotein folding: disulfide bond formation and association with calnexin. Journal of virology. 1996;70(2):778-86.
10. Yang Q, Kelkar A, Sriram A, Hombu R, Hughes TA, Neelamegham S. Role for N-glycans and calnexin-calreticulin chaperones in SARS-CoV-2 Spike maturation and viral infectivity. Science advances. 2022;8(38):eabq8678.
11. Hunegnaw R, Vassylyeva M, Dubrovsky L, Pushkarsky T, Sviridov D, Anashkina AA, et al. Interaction Between HIV-1 Nef and Calnexin: From Modeling to Small Molecule Inhibitors Reversing HIV-Induced Lipid Accumulation. Arteriosclerosis, thrombosis, and vascular biology. 2016;36(9):1758-71.
12. Alam A, Taye N, Patel S, Thube M, Mullick J, Shah VK, et al. SMAR1 favors immunosurveillance of cancer cells by modulating calnexin and MHC I expression. Neoplasia. 2019;21(10):945-62.
13. Chen Y, Ma D, Wang X, Fang J, Liu X, Song J, et al. Calnexin Impairs the Antitumor Immunity of CD4+ and CD8+ T Cells. Cancer Immunology Research. 2019;7(1):123-35.
14. Heath-Engel HM, Wang B, Shore GC. Bcl2 at the endoplasmic reticulum protects against a Bax/Bak-independent paraptosis-like cell death pathway initiated via p20Bap31. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2012;1823(2):335-47.
15. Zuppini A, Groenendyk J, Cormack LA, Shore G, Opas M, Bleackley RC, et al. Calnexin deficiency and endoplasmic reticulum stress-induced apoptosis. Biochemistry. 2002;41(8):2850-8.
16. Seto E, Yoshida M. Erasers of histone acetylation: the histone deacetylase enzymes. Cold Spring Harbor perspectives in biology. 2014;6(4):a018713.
17. Ohzono H, Hu Y, Nagira K, Kanaya H, Okubo N, Olmer M, et al. Targeting FoxO transcription factors with HDAC inhibitors for the treatment of osteoarthritis. 2023;82(2):262-71.
18. Zheng HY. Evaluating the effect of nurses' supportive and educational care on GATA2 gene expression and quality of life in patients with endometriosis. Cellular and molecular biology (Noisy-le-Grand, France). 2022;68(8):145-50.
19. Rampalli S, Pavithra L, Bhatt A, Kundu TK, Chattopadhyay S. Tumor suppressor SMAR1 mediates cyclin D1 repression by recruitment of the SIN3/histone deacetylase 1 complex. Molecular and cellular biology. 2005;25(19):8415-29.
20. Nakka KK, Chaudhary N, Joshi S, Bhat J, Singh K, Chatterjee S, et al. Nuclear matrix-associated protein SMAR1 regulates alternative splicing via HDAC6-mediated deacetylation of Sam68. Proceedings of the National Academy of Sciences. 2015;112(26):E3374-E83.
21. Pasquet M, Bellanné-Chantelot C, Tavitian S, Prade N, Beaupain B, Larochelle O, et al. High frequency of GATA2 mutations in patients with mild chronic neutropenia evolving to MonoMac syndrome, myelodysplasia, and acute myeloid leukemia. Blood. 2013;121(5):822-9.
22. Zhang LJ, Yan C, Schouteden S, Ma XJ, Zhao D, Peters T, et al. The Impact of Integrin β2 on Granulocyte/Macrophage Progenitor Proliferation. Stem cells (Dayton, Ohio). 2019;37(3):430-40.
23. Jalota A, Singh K, Pavithra L, Kaul-Ghanekar R, Jameel S, Chattopadhyay S. Tumor suppressor SMAR1 activates and stabilizes p53 through its arginine-serine-rich motif. The Journal of biological chemistry. 2005;280(16):16019-29.

24. Li F, Mandal M, Barnes CJ, Vadlamudi RK, Kumar R. Growth factor regulation of the molecular chaperone calnexin. Biochemical and biophysical research communications. 2001;289(3):725-32.
25. Lakkaraju AK, van der Goot FG. Calnexin controls the STAT3-mediated transcriptional response to EGF. Molecular cell. 2013;51(3):386-96.
26. Nowakowska-Gołacka J, Czapiewska J, Sominka H, Sowa-Rogozińska N, Słomińska-Wojewódzka M. EDEM1 Regulates Amyloid Precursor Protein (APP) Metabolism and Amyloid-β Production. International journal of molecular sciences. 2021;23(1).
27. Groenendyk J, Wang WA, Robinson A, Michalak M. Calreticulin and the Heart. Cells. 2022;11(11).
28. Kaser A, Martínez-Naves E, Blumberg RS. Endoplasmic reticulum stress: implications for inflammatory bowel disease pathogenesis. Current opinion in gastroenterology. 2010;26(4):318-26.
29. Xue Y, Meehan B, Macdonald E, Venneti S, Wang XQD, Witkowski L, et al. CDK4/6 inhibitors target SMARCA4-determined cyclin D1 deficiency in hypercalcemic small cell carcinoma of the ovary. Nature communications. 2019;10(1):558.
30. Sui H, Hao M, Chang W, Imamichi T. The Role of Ku70 as a Cytosolic DNA Sensor in Innate Immunity and Beyond. Frontiers in cellular and infection microbiology. 2021;11:761983.
31. Hu K, Babapoor-Farrokhran S, Rodrigues M, Deshpande M, Puchner B, Kashiwabuchi F, et al. Hypoxia-inducible factor 1 upregulation of both VEGF and ANGPTL4 is required to promote the angiogenic phenotype in uveal melanoma. Oncotarget. 2016;7(7):7816-28.
32. Malonia SK, Yadav B, Sinha S, Lazennec G, Chattopadhyay S. Chromatin remodeling protein SMAR1 regulates NF-κB dependent Interleukin-8 transcription in breast cancer. The international journal of biochemistry & cell biology. 2014;55:220-6.
33. Robertson KD, Ait-Si-Ali S, Yokochi T, Wade PA, Jones PL, Wolffe AP. DNMT1 forms a complex with Rb, E2F1 and HDAC1 and represses transcription from E2F-responsive promoters. Nature genetics. 2000;25(3):338-42.
34. Nebbioso A, Carafa V, Conte M, Tambaro FP, Abbondanza C, Martens J, et al. c-Myc Modulation and Acetylation Is a Key HDAC Inhibitor Target in Cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2017;23(10):2542-55.
35. Witt AE, Lee CW, Lee TI, Azzam DJ, Wang B, Caslini C, et al. Identification of a cancer stem cell-specific function for the histone deacetylases, HDAC1 and HDAC7, in breast and ovarian cancer. Oncogene. 2017;36(12):1707-20.
36. Tiffon C, Adams J, van der Fits L, Wen S, Townsend P, Ganesan A, et al. The histone deacetylase inhibitors vorinostat and romidepsin downmodulate IL-10 expression in cutaneous T-cell lymphoma cells. British journal of pharmacology. 2011;162(7):1590-602.
37. Yang Y, Yan Y, Chen Z, Hu J, Wang K, Tang N, et al. Histone Deacetylase Inhibitors Romidepsin and Vorinostat Promote Hepatitis B Virus Replication by Inducing Cell Cycle Arrest. Journal of clinical and translational hepatology. 2021;9(2):160-8.
38. Hasegawa A, Hayasaka Y, Morita M, Takenaka Y, Hosaka Y, Hirano I, et al. Heterozygous variants in GATA2 contribute to DCML deficiency in mice by disrupting tandem protein binding. Communications biology. 2022;5(1):376.
39. Rodrigues NP, Tipping AJ, Wang Z, Enver T. GATA-2 mediated regulation of normal hematopoietic stem/progenitor cell function, myelodysplasia and myeloid leukemia. The international journal of biochemistry & cell biology. 2012;44(3):457-60.
40. de Pooter RF, Schmitt TM, de la Pompa JL, Fujiwara Y, Orkin SH, Zúñiga-Pflücker JC. Notch signaling requires GATA-2 to inhibit myelopoiesis from embryonic stem cells and primary hemopoietic progenitors. Journal of immunology (Baltimore, Md : 1950). 2006;176(9):5267-75.
41. Gaine ME, Sharpe DJ, Smith JS, Colyer HAA, Hodges VM, Lappin TR, et al. GATA2 regulates the erythropoietin receptor in t(12;21) ALL. Oncotarget. 2017;8(39):66061-74.
42. Harada N, Hasegawa A, Hirano I, Yamamoto M, Shimizu R. GATA2 hypomorphism induces chronic myelomonocytic leukemia in mice. Cancer science. 2019;110(4):1183-93.
43. Brown AL, Hahn CN, Scott HS. Secondary leukemia in patients with germline transcription factor mutations (RUNX1, GATA2, CEBPA). Blood. 2020;136(1):24-35.
44. Hayakawa F, Towatari M, Ozawa Y, Tomita A, Privalsky ML, Saito H. Functional regulation of GATA-2 by acetylation. Journal of leukocyte biology. 2004;75(3):529-40.
45. Bresnick EH, Jung MM, Katsumura KR. Human GATA2 mutations and hematologic disease: how many paths to pathogenesis? Blood advances. 2020;4(18):4584-92.
46. McReynolds LJ, Yang Y, Yuen Wong H, Tang J, Zhang Y, Mulé MP, et al. MDS-associated mutations in germline GATA2 mutated patients with hematologic manifestations. Leukemia research. 2019;76:70-5.
47. Kozyra EJ, Pastor VB, Lefkopoulos S, Sahoo SS, Busch H, Voss RK, et al. Synonymous GATA2 mutations result in selective loss of mutated RNA and are common in patients with GATA2 deficiency. Leukemia. 2020;34(10):2673-87.

48. Wang H, Cui B, Sun H, Zhang F, Rao J, Wang R, et al. Aberrant GATA2 Activation in Pediatric B-Cell Acute Lymphoblastic Leukemia. Frontiers in pediatrics. 2021;9:795529.
49. Camargo JF, Lobo SA, Hsu AP, Zerbe CS, Wormser GP, Holland SM. MonoMAC syndrome in a patient with a GATA2 mutation: case report and review of the literature. Clinical infectious diseases : an official publication of the Infectious Diseases Society of America. 2013;57(5):697-9.
50. Shearstone JR, Golonzhka O, Chonkar A, Tamang D, van Duzer JH, Jones SS, et al. Chemical Inhibition of Histone Deacetylases 1 and 2 Induces Fetal Hemoglobin through Activation of GATA2. PloS one. 2016;11(4):e0153767.
51. Walsh JC, DeKoter RP, Lee H-J, Smith ED, Lancki DW, Gurish MF, et al. Cooperative and Antagonistic Interplay between PU.1 and GATA-2 in the Specification of Myeloid Cell Fates. Immunity. 2002;17(5):665-76.
52. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic acids research. 2023;51(D1):D523-d31.
53. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols. 2010;5(4):725-38.
54. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC bioinformatics. 2008;9:40.
55. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic acids research. 2007;35(10):3375-82.
56. Hansmann UHE, Okamoto Y. New Monte Carlo algorithms for protein folding. Current Opinion in Structural Biology. 1999;9(2):177-83.
57. Brooks BR, Brooks CL, 3rd, Mackerell AD, Jr., Nilsson L, Petrella RJ, Roux B, et al. CHARMM: the biomolecular simulation program. Journal of computational chemistry. 2009;30(10):1545-614.
58. Yang X-S. 3 - Optimization algorithms. In: Yang X-S, editor. Introduction to Algorithms for Data Mining and Machine Learning: Academic Press; 2019. p. 45-65.
59. Tsuruta S, Misztal I, Strandén I. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. Journal of animal science. 2001;79(5):1166-72.
60. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. Proteins. 2007;66(4):778-95.
61. Handl J, Knowles J, Vernon R, Baker D, Lovell SC. The dual role of fragments in fragment-assembly methods for de novo protein structure prediction. Proteins. 2012;80(2):490-504.
62. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic acids research. 2019;47(D1):D464-d74.
63. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 1994;22(22):4673-80.
64. Feng DF, Doolittle RF. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. Journal of molecular evolution. 1987;25(4):351-60.
65. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. Trends in biochemical sciences. 1998;23(10):403-5.
66. Ma J, Wang S. Chapter Five - Algorithms, Applications, and Challenges of Protein Structure Alignment. In: Donev R, editor. Advances in Protein Chemistry and Structural Biology. 94: Academic Press; 2014. p. 121-75.
67. Yang Y, Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, et al. SPIDER2: A Package to Predict Secondary Structure, Accessible Surface Area, and Main-Chain Torsional Angles by Deep Neural Networks. Methods in molecular biology (Clifton, NJ). 2017;1484:55-63.
68. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of computational chemistry. 2012;33(3):259-67.
69. Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein science : a publication of the Protein Society. 2018;27(1):293-315.
70. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic acids research. 2007;35(Web Server issue):W375-83.
71. Jacobson MP, Friesner RA, Xiang Z, Honig B. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. Journal of Molecular Biology. 2002;320(3):597-608.
72. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics (Oxford, England). 2014;30(12):1771-3.
73. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins. 2003;52(1):80-7.
74. Rigsby RE, Parker AB. Using the PyMOL application to reinforce visual understanding of protein structure. Biochemistry and molecular biology education : a bimonthly publication of the International Union of Biochemistry and Molecular Biology. 2016;44(5):433-7.

75. McRee DE. 3 - COMPUTATIONAL TECHNIQUES. In: McRee DE, editor. Practical Protein Crystallography (Second Edition). San Diego: Academic Press; 1999. p. 91-cp1.
76. Krengel U, Imberty A. Chapter 2 - Crystallography and Lectin Structure Database. In: Nilsson CL, editor. Lectins. Amsterdam: Elsevier Science B.V.; 2007. p. 15-50.