UNIVERSITAT DE BARCELONA

Targeting the TGF-β pathway, SMAD proteins and cofactors

Carles Torner Blancafort

Targeting the TGF-β pathway, SMAD proteins and cofactors

Carles Torner Blancafort

2024

UNIVERSITAT DE BARCELONA

Facultat de Biologia

Programa de Doctorat en Biomedicina, Universitat de Barcelona

Targeting the TGF-ß pathway, SMAD proteins and cofactors

Memòria presentada per Carles Torner Blancafort

Per optar al grau de Doctor per la Universitat de Barcelona

Realitzat a

Structural Characterization of Macromolecular Assemblies laboratory, Mechanisms of Disease Programme, IRB Barcelona

Biorecognition Laboratory, Department of Biomedicine, University of Bergen

Doctorand: Carles Torner Blancafort

Directora
Maria J. Macias Hernández

Co-Directora
Aurora Martinez

Tutor
Manuel Palacín Prieto

ACKNOWLEDGEMENTS

First, I would like to thank Maria Macias for her work as thesis director, supervisor and group leader during all these years and for giving me the opportunity to do my PhD in her laboratory at the IRB Barcelona. Thank you for your time, patience, support, and willingness to listen and discuss the ideas and topics that made this project possible. Also for helping me apply for my fellowship and also for grants to visit other labs. I am grateful for the opportunity to work in a project that combines basic and applied research, related to the fascinating TGF-β signaling pathway. It was something I had dreamed of doing, but it didn't always seem possible. I hope that one day our results will be meaningful for patients!

I would like to thank my thesis co-director, Aurora Martinez. The small molecule screening against SMAD4 would not have been possible without her tenacity and continuous help. Aurora taught me a lot about drug screening, new methods that we did not use in Barcelona, a different way of perceiving research than I was used to. Thanks to the opportunity that both Aurora and Maria gave me to enjoy a short period as a visiting researcher at the University of Bergen, I discovered a fantastic lab and incredible people. Thank you, Aurora, for all the moments, your kindness, your time and your support.

I would like to thank all those who have accompanied me over the years. Without your past and present work, none of these results would have been possible.

Lidia, thank you for giving me my first lessons in the lab in protein cloning, purification and single-molecule biophysical assays. I have never seen anyone successfully run so many protocols and supervise three master students at the same time. Thank you also for all the things you are still teaching me, all the help with protein production for SMAD4 related projects and all the moments in and out of the lab.

Thank you, Pau, for all the knowledge, talks about science and your patience for teaching me so much about computer science, tricks, analysis of clinical and chip-seq data and basically how to use a computer properly. Thanks also for the custom software, without it our lives would have been much more difficult!

Eric, thanks to your ability to produce difficult proteins and your skills in protein cloning, The RREB1 project (and many others) would not be possible without your golden hands and intuition.

To Radek, thank you for all your talks, your time, your vision and all the work we have done together on different projects. Thank you for showing me your skills in protein and protein-DNA complex crystallization, your work on RREB1-DNA structures and so much more! You are the post-doc that every PhD student needs.

To Miriam, I am grateful for your help and talent focused on understanding SMAD4 structural changes in disease variants. Your thesis will be incredible, I am sure.

To the rest of my current lab mates, Ingrid, Marina, Rameez and Rebeca, many thanks for your kindness, conversations and help! And to the ones that already left the lab, especially Tiago and Jorge, for their friendship and the incredible moments in the lab. I will never forget the “karate lessons” and “the tap dance sessions” with you.

I would also like to take this opportunity to thank all the people in the research units and platforms who have been helpful in any step of the project, as well as all the incredible people I have had the opportunity to meet during this journey, from whom I have been lucky enough to learn and have interesting scientific conversations over a nice cup of coffee or tea. In this sense, I would like to start with some researchers in Barcelona, Marta Taulés from the Unitat Anàlisi Biomolecular of the CCiTUB and Israel Ramos and Maria Caballero from the Drug Screening Unit of the IRB Barcelona. Also to all people who help me during my various visits to other European institutes, research organizations or universities through programmes such as EU-OPENSCREEN, INSTRUCT or MOSBRI. The whole team of the Aurora Martinez lab at UiB for their welcome and help, especially Kunwar Jung-KC and Emil Hausvik who were involved in the SMAD4 HTS campaign; Stephan Niebling, Angelica Struve, Osvaldo Burastero and Maria Marta Garcia at the SPC facility at EMBL Hamburg; Dean Derbyshire at the ProLinC facility (University of Linköping) and Anne-Sophie Humm, José A. Marquez and the rest of his team at the HTX lab of the EMBL Grenoble.

I would like to express my gratitude to the members of the Thesis Advisory Committee, who always gave us so much good advice and helped me progress in each evaluation. Eugenio Vázquez, Josan Márquez, Manuel Palacín and Susana de la Luna, thank you! I would also like to thank the thesis committee, Drs, Maribel Loza, Concepción Civera, and Francesc Ventura as well as Ekaitz Errasti, Miriam Royo and Jorge Cuellar for accepting to participate in this defense.

To Nora Pibernat, thank you very much for the art behind the cover of this thesis! No one could do it better than you!

I also would like to thanks to all my professors, from the elementary school to the university. Your job is key to instruct and inspire the new generations to come.

I will also like to thank the financial support (mentioned in the Annex D) received from different agencies that have made this project possible.

Last and not the least, thanks to all my loved ones, without you this journey would have been much more difficult. Als meus amics i companys de tota la vida de la Garriga i la Universitat de Girona, a la meva familia i a la Mariona, gràcies a tots de tot cor!

ABSTRACT

TGF-β signaling is key for many biological processes as embryo development, tissue homeostasis and immune system regulation. When altered, this pathway can lead to diseases such as cancer, fibrosis and rare syndromes. Key elements of the pathway are the SMAD family of transcription factors, which translate the extracellular signal received by the TGF-β receptor to the nucleus for regulation of gene expression. SMAD proteins have a characteristic structure which is shaped by an MH1 domain, for specific DNA recognition, a flexible linker region, and their MH2 domain, which can form complexes with other SMAD proteins and co-factors. This last domain is often mutated in disease, especially in the case of SMAD4 for which single point mutations and deletions have been identified in the literature. In this work, I focused on the study of SMAD4 variants associated with diseases, such as cancer, Juvenile Polyposis Syndrome, Hemorrhagic Hereditary Telangiectasia and Myhre Syndrome. With this aim, we produced different recombinant protein constructs to study the effect of these variants in their fold and binding properties. Firstly, I started with the characterization of the variants R496C- and I500V/M/T- SMAD4, associated with Myhre Syndrome. This is a gain-of-function disease that begins during embryonic development, and the alterations observed lead to the dysfunction of multiple organs. We could confirm that these specific SMAD4 variants had increased levels of SMAD4 protein in cells, possibly related to decreased ubiquitination and degradation of the protein, among other possible causes are loss-of-function variants, as in gastrointestinal cancers and Juvenile Polyposis. In this case, our work showed that the complexes with R-SMADs and the variants lead to several different stoichiometries compared to those of the wild type (WT) protein.

The second section of this thesis is focused on the search for small-molecules as SMAD4 binders. We used single molecule biophysics and structural biology to identify pharmacological strategies based on targeting SMAD4 to modulate TGF-β signaling. This search was conducted through a target-based in vitro approach using purified SMAD4 MH2 domain and large libraries of compounds. Among these compounds, we included FDA-approved drugs in case we could identify hits that could be repurposed to treat individuals suffering from very rare syndromes. Validated hits have affinities of interaction ranging between low and high micromolar and will be further developed and tested. Some interesting. approved drugs were identified as SMAD4 binders.

In the last chapter of this project, I focused on the DNA recognition ability of Ras Responsive Element Binder 1 (RREB1). RREB1 plays a key role in communication between RAS and TGF-β signaling to regulate epithelial-to-mesenchymal transition

(EMT) during embryonic development and maintenance of healthy tissue, but also during cancer progression. RREB1 is a zinc finger (ZF) protein with multiple isoforms. In particular, I studied a well-conserved evolutionary ZF pair located at the C-terminus of the protein.

INDEX

LIST OF TABLES AND FIGURES

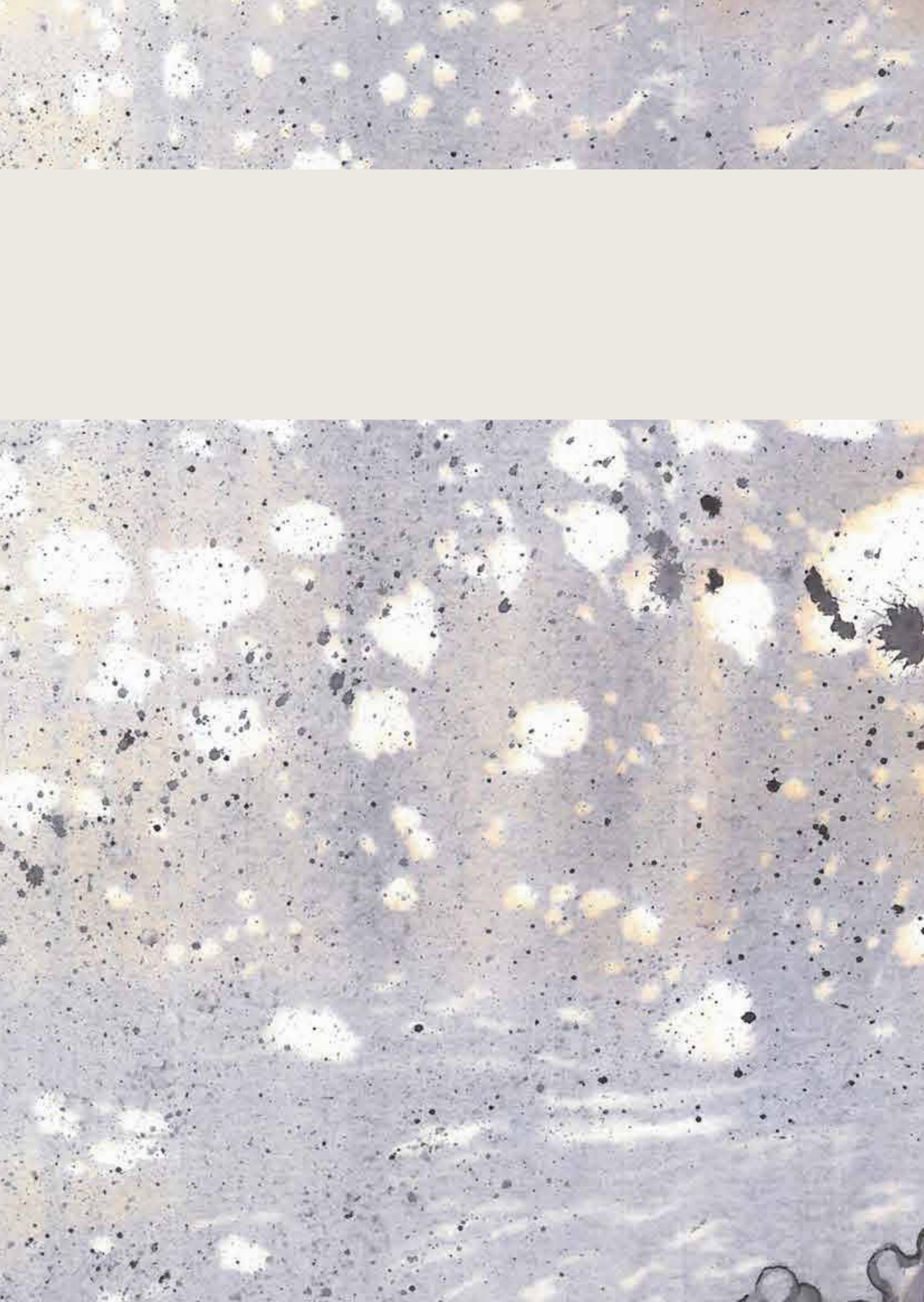
LIST OF ABBREVIATIONS

3C	Human rhinovirus 3c protease
Å	Angstrom
AMH	Anti-mullerian hormone
BMP	Bone morphogenetic protein
BMPR	BMP Receptor
CD	Crystal Direct
CDK	Cyclin-dependent kinases CDK
Co-SMAD	Common partner SMAD
CPA	Cryoprotective agent/Cryoprotectant
CRIMS	Crystallization Information Management System
D_{max}	Maximum distance
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DPC4	Deleted in Pancreatic Carcinoma locus 4
DRA	Dose-response assay
dsDNA	Double-stranded DNA
DSF	Differential Scanning Fluorimetry
ECBD	European Chemical Biology Database
ECM	Extracellular matrix
E. coli	*Escherichia coli*
EMA	European Medicines Agency
EMBL	European Molecular Biology Laboratory
EMSA	Electrophoretic shift assay
EMT	Epithelial to mesenchymal transition
ESCC	Esophageal Squamous Cell Carcinoma
FDA	U.S. Food and Drug Administration
FPLC	Fast protein liquid chromatography

FL	Full-length
GDF	Growth differentiation factor
GSK	Glycogen synthase kinase
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HHT	Hemorrhagic Hereditary Telangiectasia
HPLC	High-performance liquid chromatography
HSCs	Hepatic Stellate Cells
IPTG	Isopropyl β- d-1-thiogalactopyranoside
I-SMAD	Inhibitory SMAD
ITC	Isothermal titration calorimetry
JPS	Juvenile Polyposis Syndrome
K_D	Dissociation constant
kDa	Kilo Dalton
k_{off}	Dissociation rate constant
k_{on}	Association rate constant
LAP	Latency-associated peptide
LB	Luria Broth
MALS	Multiangle light scattering
MES	2-(N-morpholino) ethanesulfonic acid
µM	Micromolar
mM	Millimolar
nm	Nanometer
MP	Mass photometry
MyS	Myhre Syndrome
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PEG	Polyethylene glycol
POI	Protein-of-interest

PPI	Protein protein interaction
ProLinC	PROtein folding and Ligand INteraction Core facility
PY	Proline-Tyrosine motif
RAS	Rat sarcoma virus
RRE	Ras Responsive Element
RREB1	Ras Responsive Element Binder 1
Rg	Radius of gyration
R-SMAD	Receptor regulated SMAD
SAD	SMAD activation domain
SARA	SMAD anchor for receptor activation
SAXS	Small angle x-ray scattering
SBE	SMAD-binding element
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SEC	Size-exclusion chromatography
SMAD	Small mother against decapentaplegic
SMAD4 SAD-MH2	Extended MH2 domain at the N-terminus
SMAD3 MH2EEE	Phosphorylation mimic mutation 1
SMAD3 MH2 DVD	Phosphorylation mimic mutation 2
SOC	Super optimal broth with catabolite repression
SPC	Sample Preparation and Characterization Facility
SPR	Surface Plasmon Resonance
SUMO	Small ubiquitin-like modifier
$t_{1/2}$	Complex half-life time
TCEP	Tris(2-carboxyethyl)phosphine
TEV	Tobacco etch virus
TGF-β	Transforming growth factor beta
TGF-βR	Transforming growth factor beta receptor
T_m	Melting temperature

TRIC	Temperature-Related Intensity Change
TRIS	Tris(hydroxymethyl)aminomethane
TSP1	Thrombospondin 1
ULP1	Ubl-specific protease 1
WT	Wild type
ZF	Zinc Finger



INTRODUCTION

1.INTRODUCTION

1.1 General overview of the work

Since records have been kept (and even before), from ancient communities to the present day, people have sought to relieve pain, prevent and treat infections, and alleviate and cure the symptoms of disease.

Nowadays, therapeutic options have evolved to the point where we have numerous tools to treat a wide range of diseases. These tools include drugs of synthetic and natural origin, antibodies, and other biologics, as well as novel therapies, including the use of CRISPR technology and gene-editing tools to modify genes with harmful mutations, CAR-T cell therapy, new vaccines, and many others that are revolutionizing the field of medicine every day.

There are still unmet medical needs that require the identification of new -or complementary- treatments at a cost that is not prohibitive for the public health system. The identification of new compounds with pharmacological applications typically requires time and a substantial investment. However, if successful, large-scale production of chemical compounds could be less expensive than other alternatives, offsetting the initial economic investment, and facilitating its commercial production and use worldwide at an affordable cost. These needs include finding new treatments for cancer patients who have developed resistance to approved drugs and are running out of pharmacological options, or for individuals suffering from rare diseases, to name just a few. Rare diseases, in fact, are often overlooked by pharmaceutical companies because of the small number of people affected, the limited knowledge of the disease, and the lack of correlation between the observed phenotypes and the molecular basis.

With this in mind, we set out to explore the possibility of identifying molecules that could modulate the TGF-β signaling pathway, one of the seven signaling pathways conserved across metazoan, combining the expertise of our lab at the IRB Barcelona, led by the ICREA research Prof. Maria J. Macias, with this biological system, and that of Prof. Aurora Martinez, at the University of Bergen, related to the identification of small compounds with pharmacological activity. To achieve this aim, we applied several molecular biology tools and complementary biophysical techniques. TGF-β signaling, in brief, includes a family of cytokines and membrane receptors that respond to these cytokines and, in the canonical pathway, a family of transcription factor proteins that act as the messengers of the receptor signals in the nucleus. This family of proteins are known as SMAD (Mothers against Decapentaplegic) proteins **(Attisano *et al.*, 1993;**

Wrana *et al.*, 1994; Feng and Derynck, 2005; Massagué, Seoane and Wotton, 2005). SMAD-driven signaling is involved in many essential aspects of metazoans life, including embryo development or cell homeostasis **(Huminiecki *et al.*, 2009; Massagué, 2012)**. Because of its importance to the proper functioning of our cells, this signaling network is tightly regulated. Unfortunately, this signaling network is not error-free, and mutations in SMAD proteins, particularly within SMAD4, have been associated with human diseases such as cancer and rare diseases **(Massagué and Sheppard, 2023)**.

SMADs are composed of an N-terminal domain that interacts with DNA, a linker, and a C-terminal domain that participates in protein-protein interactions (PPIs) **(Shi and Massagué, 2003; Macias, Martin-Malpartida and Massagué, 2015)**. Both of these domains are unique to SMAD proteins. Another characteristic of SMAD proteins is to associate among them to form heterotrimers, which is the core transcriptional unit. The functional capabilities of the core SMAD complex are further modulated by the formation of SMAD complexes with other proteins (co-activators and repressors, ubiquitin ligases, kinases, phosphatases, and chromatin remodelers, to name a few) that fine-tune the functional properties of the SMAD-driven signaling system according to cellular needs **(Fuentealba *et al.*, 2007; Sapkota *et al.*, 2007; Alarcón *et al.*, 2009; Aragón *et al.*, 2011)**.

While major therapeutic strategies to tackle TGF-β pathway are focusing on modulating the membrane receptor function or inhibiting the hormone activation **(Attisano et al., 1993; Akhurst, 2017; Cho *et al.*, 2020; Liu, Ren and Ten Dijke, 2021; Yap *et al.*, 2021; Shi *et al.*, 2022)**, no therapeutic strategies have been tested in preclinical or clinical assays targeting SMAD proteins. Targeting SMAD4 can be of special interest since it is the most mutated element in the SMAD driven TGF-β pathway in primary tumors, especially in pancreatic and gastrointestinal tract cancers, and has key roles in advanced cancer stages, fibrosis and rare diseases. Individuals with Juvenile Polyposis Syndrome (JPS) or Hereditary Hemorrhagic Telangiectasia (HHT) **(Miyaki and Kuroki, 2003; Cao, Plazzer and Macrae, 2023)** usually have alterations in the proper function of epithelial tissue in various organs. The SMAD4 variants associated with these epithelial disorders, which accumulate mainly in the MH2 domain of the protein, cause inhibition of SMAD complex formation. Individuals with Myhre syndrome (MyS) have specific SMAD4 point mutations associated with stabilization of SMAD proteins. Remarkably, variants linked to rare diseases are often found as well in cancer patients.

In addition to these applied aims, we also planned to contribute to a better understanding of the molecular mechanisms of Epithelial to Mesenchymal transition (EMT), a phenotypic characteristic required during development and tissue repair but that can promote cancer invasion and metastasis in scenarios associated with disease **(Nieto, 2011)**. EMTs are driven by specialized signaling events that activate the expression of a set of transcription factors (**EMT TFs**) that repress epithelial genes and induce the expression of mesenchymal features **(Batlle *et al.*, 2000; Cano *et al.*, 2000)**. For our studies, we have selected a specialized effector of RAS/MAPK signaling, RREB1 (RAS response element binding protein 1) that also receives inputs from TGF-β to induce EMT and metastatic outgrowth in carcinoma cells **(Janda *et al.*, 2002; David *et al.*, 2016; Deng *et al.*, 2020; Su *et al.*, 2020)**. RREB1 is a large multi-Zinc finger (abbreviated as ZF) protein, four times longer than average protein sequences in eukaryotes **(Brocchieri and Karlin, 2005)**. The ZF domains are the most abundant DNA binding structures found in eukaryotic transcription factors, present in more than 800 proteins in the human proteome **(Wolfe, Nekludova and Pabo, 2000; Najafabadi *et al.*, 2015)**. The ZFs of RREB1 are grouped into three main clusters, separated by large intervening regions lacking other known structured domains. Our contribution in this PhD thesis has been to analyze the interactions between the cluster of ZFs located at the C-terminal part of the protein and specific DNA motifs. This project is carried out as a collaboration with the laboratory of Dr. Joan Massagué (Cancer Biology and Genetics Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA).

1.1.1 Working hypotheses of this work

Many SMAD variants associated with cancer are thought to correlate with a loss-of-function role of SMAD4 proteins **(Miyaki and Kuroki, 2003; Chacko *et al.*, 2004; Massagué and Sheppard, 2023)**, whereas in Myhre syndrome (MyS), a rare disease affecting embryo development and multiple organs, these mutations are correlated with a gain of function which lead to increased SMAD4 protein levels and decreased ubiquitination of the protein in patient cell lines **(Le Goff *et al.*, 2011; Caputo *et al.*, 2014)**. Thus, as SMAD proteins form quaternary structures, we have hypothesized that some of these MyS mutations and cancer variants might affect the stoichiometry of the SMAD complexes, giving rise to transcriptional complexes of modified selectivity and affinity for DNAs and cofactors. We also hypothesized that these mutations might affect the stability of these SMAD heterocomplexes, modifying the duration of the transcription activation of specific genes and giving rise to diseases.

Driven by the urgent societal need to find new treatments for cancer patients and also for individuals with Myhre syndrome and other rare diseases, we proposed the transcription factor SMAD4 as a target for drug discovery. If we could find small molecules that interact with SMAD4, these molecules could be developed either as research tools or as molecules with potential pharmaceutical application, depending on their specific action. We also planned to test drugs already on the market for drug repurposing, an option that will allow us a faster path to the clinic if effective compounds are found, avoiding the need for time- and cost-demanding toxicity studies.

1.1.2 Outline of results

The results section is distributed in three chapters. The first one is focused on studying the quaternary structures of SMAD proteins and how a few selected disease-associated mutations in SMAD4 affect the tertiary and quaternary structure of SMAD complexes. The second chapter has been dedicated to identifying compounds that interact with SMAD4 to modulate SMAD interactions affected in disease associated variants. The third chapter includes our studies of the C-terminal region of RREB1 and DNA.

The grants that have supported this research are mentioned in the Annex D.

1.2 A brief overview of TGF-β receptors and cytokines

Inter and intra cellular communication and signaling are key for the regulation of almost all processes in human cells. TGF-β family of cytokines are key players in many processes such as embryo development, differentiation, homeostasis, and immune system surveillance, to mention a few. The context-dependent action of TGF-β signaling determines cell fate in health and disease, as it plays key roles in cancer, fibrosis, chronic inflammation and congenital skeletal, connective or cardio-vascular diseases **(Massagué, 2012; Massagué and Sheppard, 2023)**.

TGF-β response programs vary depending on the cellular context and tissues and on the intercommunication with other essential signaling pathways. In epithelial and endothelial cells, TGF-β dictates the phenotypic transition of cell groups, their differentiation and paracrine secretion. One of the key mechanisms regulated in this group of cells is epithelial to mesenchymal transition (EMT) necessary during development and apoptosis but also promoting cell migration and invasion of distal tissues in advanced cancers, among others. Mesenchymal cells (fibroblasts, bone and connective tissue) are also regulated by this signaling pathway, which is determinant in key roles such as regulation of extracellular matrix (ECM) production or cell migration and motility. Modulation of migration is also reported in cells from the nervous system, which also needs TGF-β input for survival **(Kashima and Hata, 2018)**. Immune cell systems, innate and adaptive, are also regulated by this pathway, observing some differences between cell subtypes. All these processes are possible thanks to the capability of TGF-β to determine the transcriptional landscape of cells **(Wang *et al.*, 2022)**.

Although we use the term TGF-β generically, the TGF-β family encompasses around 40 secreted cytokines, classified in subfamilies based on their structure and biological role in cell signaling. One is the TGF-β/Nodal subfamily, which comprises TGF-β1, TGF-β2, and TGF-β3 receptors (TGF-β for short), as well as Nodal, Activin A-E, GDF1, GDF3, GDF8/Myostatin, GDF9 and GDF11. Inhibin, Lefty1 and Lefty2 are also members of this subfamily, although they function as inhibitors of Activin receptors or Nodal co-receptors respectively. The second subfamily of cytokines is the bone morphogenic protein (BMP) that activates the BMP-pathway. This subfamily is composed of BMP2, BMP4, BMP5, BMP6, BMP7, BMP8, BMP8B, BMP9, BMP10, BMP15, the anti-müllerian hormone (AMH) and GDF5, GDF6, GDF7 and GDF10. BMP3 functions as an inhibitor of BMP receptors **(Plouhinec, Zakin and De Robertis, 2011)**.

All three TGF-β cytokines (TGF-β1, TGF-β2, TGF-β3) are synthesized as prohormones. During maturation, the prohormone is cleaved, producing the mature hormone and a short peptide, the latency-associated peptide (LAP). Both components associate non-covalently to produce an inactive form of the cytokine (Latent TGF-β), which is exported outside the cell, where it is exposed in the cell membrane by partner proteins or retained in the extracellular matrix. Finally, the LAP undergoes conformational changes upon interaction with enzymes or other proteins specifically located on the cell membrane, that end releasing the mature and dimeric TGF-β cytokine **(Massagué, 2000; Massagué and Sheppard, 2023)**. Active cytokines interact with their specific target receptor to activate the signaling cascade **(Figure 1).**

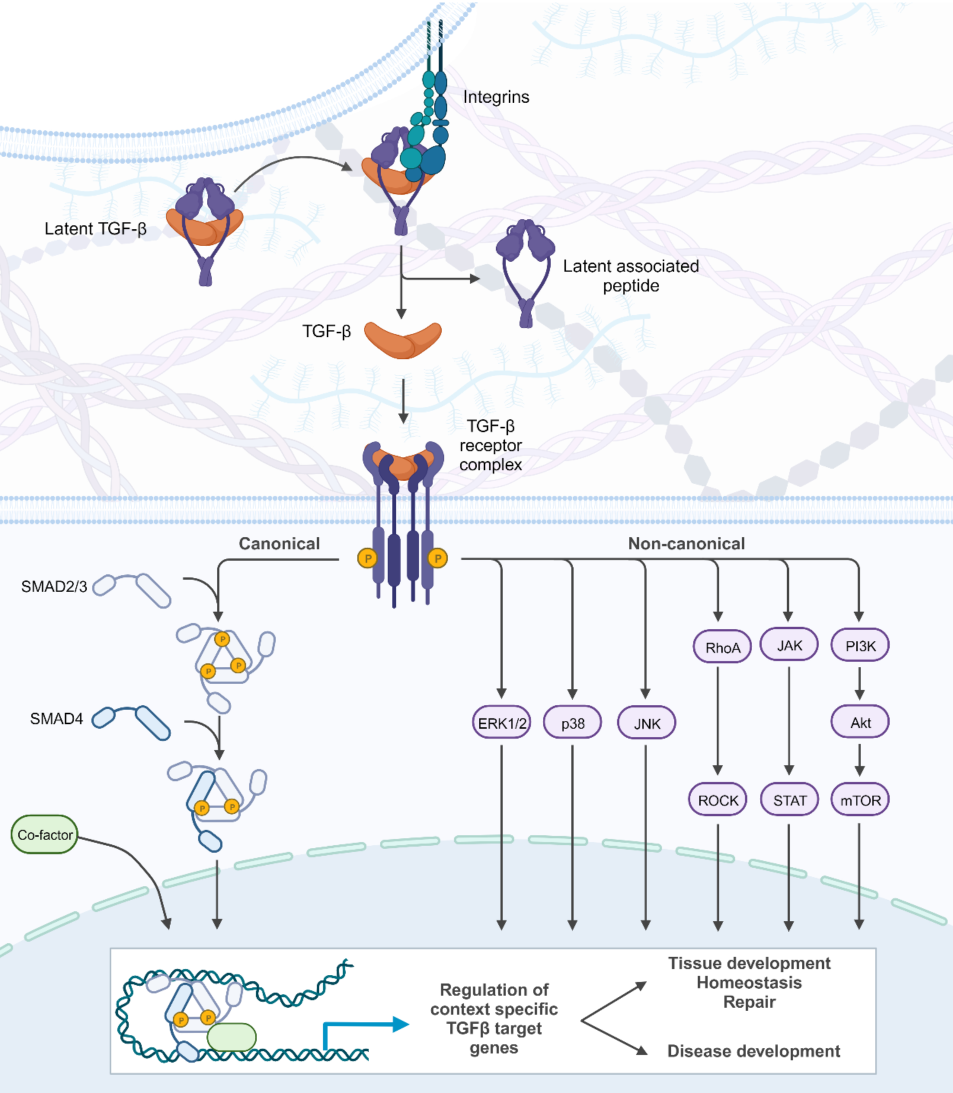


Figure 1. Schematics describing the TGF-β signaling pathway, including the TGF-β hormone, the receptor and the canonical and non-canonical routes of signaling.

Upon activation, each cytokine subtype interacts with specific cell membrane receptor systems that initiate signal transduction. In the case of TGF-β, there are different subtypes of membrane receptors known as type I and type II receptors (TGFβRI and TGFβRII for short). These receptors are composed of an extracellular domain, which binds to the hormone, and an intracellular serine/threonine kinase domain. TGF-β hormones bind to TGFβRII, and subsequently the TGFβRI associates to them to form a ternary complex formed by two subunits of each receptor subtype (assembled as a dimer of heterodimers) although the stepwise mechanism of the receptor formation is not fully understood **(Hinck, Mueller and Springer, 2016)**.

Once the hormone/receptor complex is formed, the type II receptor phosphorylates the type I receptor in the cytosol of the cell. Then, the type I receptor phosphorylates R-

SMAD proteins (TGF-β canonical pathway) or other substrates (TGF-β non-canonical pathway) to transmit the extracellular signal to the nucleus. R-SMADs are doubly phosphorylated at the C-terminus at a Ser-x-Ser motif and are activated as a result.

1.3 SMAD proteins as drivers of TGF-β signaling

SMAD proteins are transcription factors composed of two globular domains (MH1 and MH2 domains) and a flexible linker region that connect them **(Gomes *et al.*, 2021)**. Three SMAD protein classes are reported in the bibliography **(Shi and Massagué, 2003; Macias, Martin-Malpartida and Massagué, 2015)**. One class is defined by the Receptor activated SMADs (R-SMADs). In vertebrates, these proteins are SMAD2/3, which are phosphorylated by TGF-β family receptors, and SMAD1/5/8, which are activated by the BMP receptor family **(Figure 2, Table 1)**. The receptor phosphorylation site is located at the very C-termini of R-SMADs. In contrast, SMAD4 does not require receptor activation and forms complexes with activated R-SMADs for its function.

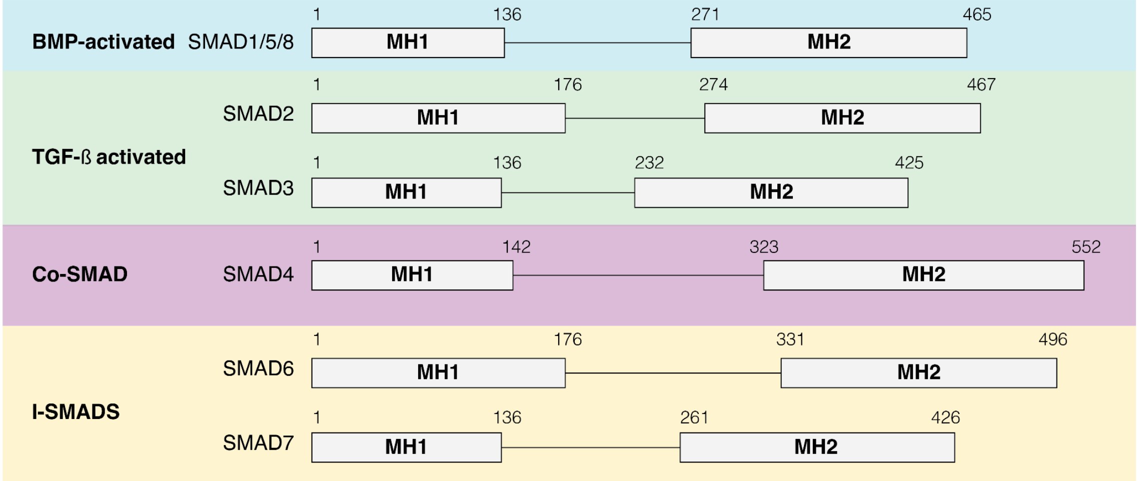


Figure 2. Schematic representation of SMAD protein members and their domains.

Since SMAD4 can associate with BMP and TGF-β/Nodal activated SMAD proteins, it is also known as the Co-SMAD. The third type of SMAD proteins are the inhibitory SMADs or I-SMADs (SMAD6/7) **(Figure 2, Table 1)**. These I-SMADs are more divergent in sequence, as they only have a well-defined MH2 domain **(Macias, Martin-Malpartida and Massagué, 2015)** that competes with R-SMADs for binding to receptors and modulators.

Table 1. Different SMAD proteins classify according to their functional role and pathway

SMAD	Receptor Activated	Common SMAD	Inhibitory SMAD	Pathway
SMAD1/5/8	✓			BMP
SMAD4		✓		BMP & TGF-β
SMAD6			✓	BMP
SMAD7			✓	TGF-β

The MH1 domain is key in specific DNA recognition through interactions with the DNA major groove. The protein-DNA interactions are mediated by a conserved β-hairpin in this domain. These interactions were characterized using a palindromic motif called SBE (5′-GTCTAGAC-3′). The SMAD-DNA interaction was structurally characterized using X-ray crystallography and also by single-molecule biophysics. Later on, our lab together with that of Dr. Joan Massagué (Sloan Kettering, New York, USA) revealed how the MH1 domain of SMAD proteins can also recognize specific GC-rich motifs (5-GC sites) in key genes and regulatory regions of TGF-β activated genes. We have also observed that SMAD protein binding with SBE and 5-GC motifs is not identical in all SMAD proteins, since BMP activated SMADs interact with these sites as dimers, whereas TGF-β activated SMADs and SMAD4 do so as monomers **(Shi *et al.*, 1998; BabuRajendran *et al.*, 2010; Baburajendran *et al.*, 2011; Martin-Malpartida *et al.*, 2017; Ruiz *et al.*, 2021)**.

Once in the nucleus, R-SMADs-SMAD4 complexes undergo two rounds of consecutive phosphorylations in the linker connecting the MH1 and MH2 domain. The first phosphorylation is carried out by cyclin-dependent kinases CDK8/9 **(Matsuura *et al.*, 2004)**, and then, by glycogen synthase kinase-3β (GSK3β) **(Fuentealba *et al.*, 2007)**. The functional outcome of these phosphorylations in SMAD complexes is different. CDK8/9 phosphorylation enhances the transcriptional activity of R-SMADs, by increasing their affinity for transcription activators as YAP1 for BMP driven signaling, and PIN1 in the case of TGF-β/Nodal activated R-SMADs **(Alarcón *et al.*, 2009; Aragón *et al.*, 2011, 2012)**. Moreover, CDK8/9 phosphorylations can be reversed by the action of specific phosphatases **(Liu and Feng, 2010)**. However, after GSK3β phosphorylation, the linker of R-SMADs becomes a binding site for the HECT family of E3 ubiquitin ligases, which marks SMADs for targeted protein degradation and this is a point of no return because it cannot be reversed by phosphatases **(Alarcón *et al.*, 2009)**. It has been surprising that the recognition of either the CDK8/9 sites and also of those generated by GSK3β is driven by proteins that have in common the presence of WW domains. These domains recognize proline rich motifs and act as protein-protein interaction modules in many signaling proteins **(Macias *et al.*, 1996, 2000; Macias, Wiesner and Sudol, 2002)**. They recognize PY and phosphorylation motifs present in R-SMADs **(Aragón *et al.*, 2011, 2012)**.

Most of the structural work has been carried out using independent domains until recently. In fact, the conformational ensemble displayed by full-length SMAD4 and SMAD2 proteins have been studied by SAXS in our lab and in collaboration with Dr. Tiago Cordeiro (NOVA University Lisbon, Portugal). Under the experimental conditions investigated in this study, the full-length SMAD4 protein behaved as a monomer, whereas SMAD2 has a high tendency to form dimers and trimers through interactions of the MH2 domains. In both proteins, it has been observed the presence of both open and closed conformations in solution **(Gomes *et al.*, 2021)**. This is important for the interpretation of the stoichiometry of functional complexes with R-SMADs and for the regulation of DNA recognition and PPIs in the cell **(Figure 3)**.

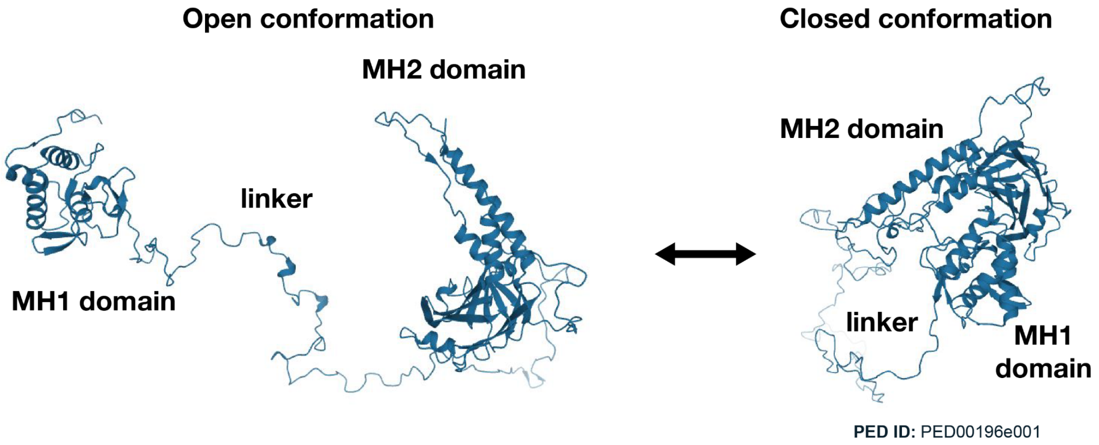


Figure 3. SMAD4 structure in solution.

Open and closed conformations are shown. The MH1 domain, linker region and MH2 domain are labeled.

1.4. Different stoichiometries in SMAD complexes according to scientific literature

SMAD signaling starts with R-SMAD activation (phosphorylation at the C-terminus), which enhances R-SMAD capacity to form homo- and hetero-oligomeric species. Studies on liver cell lines also suggest a key role of SMAD2-SMAD4 heterotrimeric complexes of different SMAD composition **(Lucarelli *et al.*, 2018)** . In recent years, other oligomeric SMAD complexes have been described with and without the presence of SMAD4. Using zebrafish embryos, an *in-vivo* model widely used to study embryonic development, it was shown that SMAD4 is essential for BMP-activated SMAD1/5 signaling. Nuclear localization of SMAD3 might occur in the absence of SMAD4, as reported by our lab and that of Dr. J. Massagué **(Aragón *et al.*, 2019)** although gene expression needs the recruitment of SMAD4-SMAD2 complexes after activation of the pathway, to give rise to SMAD4-SMAD2-SMAD3 complexes.

1.5 SMAD4 and its role in health and disease

In healthy individuals, SMAD4 plays a key role in promoting EMT during embryonic development. EMT is one of the key processes that allows cell differentiation and induces epithelial cells to undergo changes that affect their shape and cell-cell contacts, cytoskeletal organization, mobility and motility, production of ECM components, and basically their gene expression profile. At different stages, EMT can be controlled by Nodal or BMP, with SMAD4 being a key effector of both stimuli. Besides, SMAD4 knockout in mice is lethal and affects embryo development at distinct stages such as gastrulation or expression of mesodermal markers **(Sirard *et al.*, 1998)**, demonstrating its central role during such processes. EMT is also induced by TGF-β signaling and SMAD4 in adult tissues, where EMT-induced apoptosis contributes to their homeostasis.

TGF-β signaling also modulates the activity of the innate and adaptive immune systems. Increased signaling inhibits inflammation and the action of various immune cells, while a lack of TGF-β signaling can lead to uncontrolled inflammation and fibrosis. Alteration of SMAD4 and other components of the TGF-β pathway, either within the immune cells or in other cell lineages, can lead to severe alterations of the immune system.

SMAD4 is the most altered component of the canonical TGF-β pathway in cancer, as the gene is affected in 4-7% of cancer patients **(Figure 4)** and plays key roles in the initiation and advanced stages of cancer, metastasis or fibrosis (either in the lung, liver, kidney, or skin) **(Macias, Martin-Malpartida and Massagué, 2015)** and references herein. In some tumors, such as breast cancer or advanced stages of pancreatic or gastrointestinal cancers, SMAD4 has been implicated in aberrant activation of EMT processes that stimulate tumor cell motility and dissemination, angiogenesis, and metastasis. In fibrosis, SMAD4-activated signaling in immune cells is associated with increased inflammation, while in other cases, such as in hepatic stellate cells (HSCs), SMAD4 signaling promotes ECM production that triggers inflammation and immune response. Active SMAD4 WT heterotrimers may then be a potential target in such scenarios.

SMAD4 variants in cancer are reported to mainly lead to loss-of-function. This feature is also shared in rare diseases such as Juvenile Polyposis Syndrome (JPS), Hereditary

Hemorrhagic Telangiectasia (HHT), or JPS/HHT combined syndrome **(Cao, Plazzer and Macrae, 2023)**. Missense-mutations are the most abundant alterations and are distributed all along the SMAD4 sequence, although more abundant in its MH2 domain. Some of the mutations localized in the MH2 domain are located at the interface of interaction with R-SMADs, leading to the hypothesis that they can affect SMAD4-R-SMAD complexes.

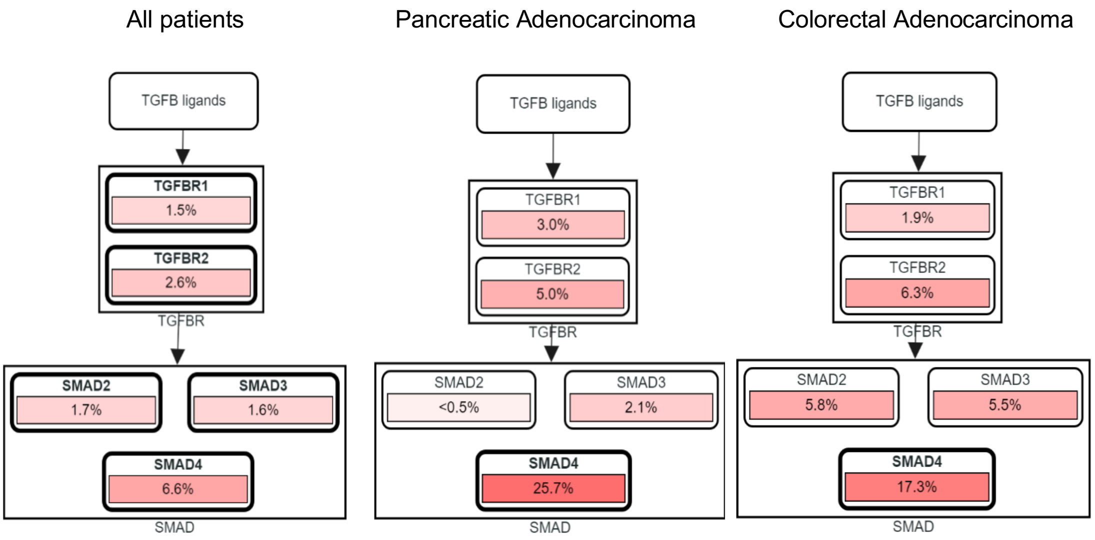


Figure 4. Alteration in the abundance of key TGF-β/Nodal signaling pathways in different cancer patient cohorts. SMAD4 is the most altered component of the signaling cascade and is highly altered in pancreatic and gastrointestinal cancers. Data was retrieved from MSK MetTropism study **(Nguyen *et al.*, 2022)** and analyzed with cBioPortal **(Cerami *et al.*, 2012; Gao *et al.*, 2013; de Bruijn *et al.*, 2023)**. The diagram was generated using PathwayMapper **(Bahceci *et al.*, 2017)**.

JPS is one of the best examples to illustrate the loss of function behavior of SMAD4 variants described in non-cancer diseases. JPS is caused by an alteration in the control of the growth of epithelial cells, typically in the colon or other areas of the digestive tract. This disease is characterized by the appearance of benign polyps in these organs, which leads to gastrointestinal bleeding, anemia, abdominal pain and diarrhea. Between 40-50% of JPS patients have germ line disease-causing variants (DCV) in SMAD4 or BMPR1A, a subtype of type 1 BMP receptor which phosphorylates SMAD1 and SMAD5. These patients also have an increased risk of gastrointestinal cancer. SMAD4 mutations accumulate specially in the MH2 domain **(Figure 5)**. Similar mutations have been reported in HHT, which is characterized by changes in the epithelial tissue of blood vessels that prevent normal blood circulation and proper connection to veins and arteries. This condition manifests as nosebleeds and bleeding in various organs, including the colon, lungs and brain, which can be life-threatening. Combined HTT/JPS is also reported in the bibliography. Patients with such clinical phenotype are reported to have alterations in SMAD4, but not in other proteins of the TGF-β signaling cascade.

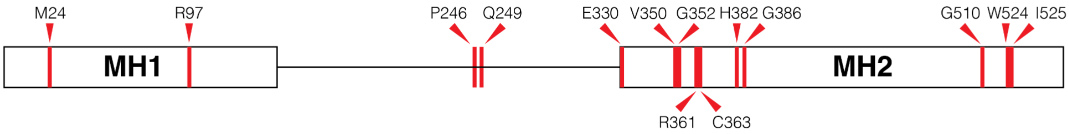


Figure 5. Non-redundant location of JPS SMAD4 variants reported in five different studies.
The position of the amino acids affected by missense mutations are indicated.

In contrast to the loss of function effect, variants at two sites of the MH2 domain, namely R496C and I500M/T/V have been described as causing a gain of function, and are key in the progression of the Myhre syndrome (MyS) **(Le Goff *et al.*, 2011; Le Goff, Michot and Cormier-Daire, 2014)**. MyS is an incurable, rare disorder affecting connective tissue. Individuals with this condition display distinct body features, heart and aorta problems, hearing loss and intellectual disability, including autistic-like behavior.

In tissues and primary derived cell lines of MyS individuals carrying described mutations, it has been observed decreased levels of ubiquitinated SMAD4, increase of SMAD4 protein levels, as well as of phospho R-SMADs and changes in transcription of genes associated with TGF-β and BMP pathway. Three additional individuals diagnosed as MyS do not show any of the characteristic variants, giving the possibility of certain variability on the gene signature that drives the condition or that these other individuals show similar phenotypes with a different molecular origin.

Since the first mutations in residue I500 were discovered in MyS, several hypotheses were proposed to explain its effects on the protein properties or activity. Some of them are changes in thermal stability of SMAD4 MH2 domain, changes in the orientation of Lys519 (a target for SMAD4 ubiquitination) or more stable SMAD4 complexes **(Le Goff *et al.*, 2011)**.

Based on this abundant knowledge regarding the SMAD complex determinants and the potential influence of SMAD4 mutations in these interactions, in this work, we set out to establish a protocol to characterize SMAD quaternary structure using molecular and biophysical approaches. All these techniques are described in the Materials and Methods section, and the results are described in the first chapter of the Results section.

1.6 Current strategies to tackle and modulate TGF-β and BMP pathways

At the writing of this thesis, there are no drugs targeting the TGF-β pathway in the market. However, different strategies to modulate TGF-β are currently in development or even in preclinical and clinical trials for a sort of different disease, especially in cancer, where the key role of the pathway got the interest of many oncologists. Different sorts of pharmacological products are proposed, which will differ according to the final therapeutic aim.

The main strategy so far has been to inhibit the signaling pathway by either stopping the production of the cytokine, blocking the cytokine-receptor interaction or inhibiting the intracellular kinase of the receptors **(Figure 6)**. There are no approaches currently described that act on effectors located downstream the receptors in the pathway, such as the SMAD proteins.

A recent approach to inhibit the pathway is the use of Bintrafusp alfa, a bifunctional fusion protein combining the extracellular domain of the TGF-β type II receptor and an IgG that blocks PD-L1, a membrane protein, and has been currently discontinued as reported in the web page of Merck. The use of small molecules to block the receptor kinases, such as Vactosertib in conjunction with gemcitabine increased the antitumor activity of gemcitabine, another approach also in clinical trials for patients with pancreatic cancer **(Lee *et al.*, 2023)**. Inhibition strategies are also proposed for fibrosis and drugs like Pirfenidone, which reduces TGF-β1 production, or NIS793, an anti-TGF-β monoclonal antibody, are being tested in Phase II/III and Phase II clinical trials.

There are also some attempts to use the cytokine BMP-2 to induce bone tissue repair and regeneration to recover from fractures **(Hustedt and Blizzard, 2014; Zamarioli *et al.*, 2022)**.

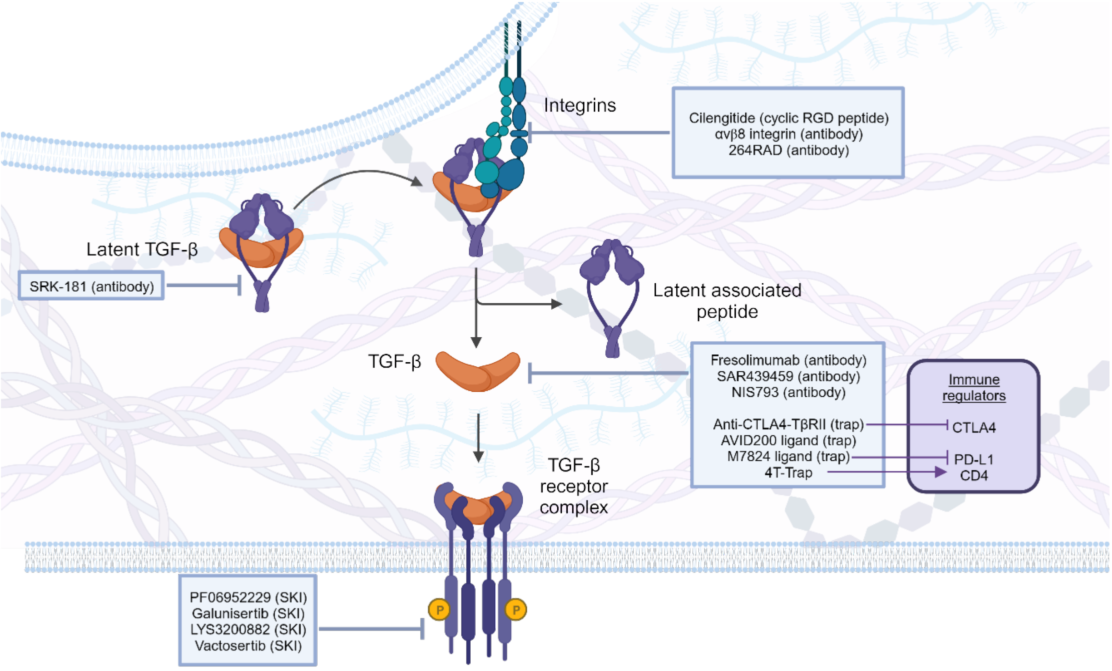


Figure 6. TGF-β pathway inhibition strategies in pre-clinical or clinical assays for cancer treatment. The type of pharmaceutical product is indicated, including antibodies, TGF-β traps, cyclic peptides and small-molecule kinase inhibitors (SKI). Diagram adapted from Liu, S et al. Signal Transduct Target Ther, 2021.

1.7 Drug discovery and screening methods using purified target protein

Early stage small molecule drug discovery campaigns usually include a screening assay to identify compound hits, and can be divided into phenotypic and target-based screening. Phenotypic screening is based on observing changes in the activity or behavior of specific cell lines, tissues, or organoid models, after treatment with compound libraries. It is a rapid approach, easy to perform as a high throughput (HTP) assay and can provide significant results in a short period of time. However, its initial design aiming target selectivity is complicated and time-consuming, and it tends to result in a high number of false positives. In addition, because it is a cellular assay, the lack of confidence in specific target engagement makes it difficult to improve the initial hits in a rational way. This lack of information further complicates the process of optimizing the progression of hits to leads, which are modified hits with improved activity, selectivity, pharmacokinetics and safety.

Target-based screening supported by structure-based drug discovery provides a convenient platform to identify molecules that bind directly to a specific biomolecule **(Tahk *et al*, 2023)**. This screening can be performed either *in silico* or experimentally, or in a combination of both. Target-based screening is an attractive approach in cases where a mutated gene has been identified as being associated with the disease.

For *in silico* approaches, the application of software packages like the Schrödinger platform, or new ones using machine learning based tools is driving a growing interest in exploring previously uncharacterized protein-protein binding sites that can be used as hotspots in library screening. However, the main drawback of these *in silico* approaches is that the predicted –usually numerous– hits need to be purchased and validated experimentally to validate that the hits indeed bind to the targets within the micromolar range of affinity.

If screening is done experimentally, it requires the availability of pure and well-behaving proteins (in mg amounts) and an efficient high throughput system for screening the compound libraries and analyzing the results. This strategy may suffer from potential off-target effects during validation in cells due to lack of specificity, and as with any approach, its success is highly case-specific.

For our work, we set to define SMAD4 as a target for modulating TGF-β signaling. In the past, pharma companies as well as many research groups put the effort in targeting the TGF-β receptor. Although the receptor is an attractive target, efficient inhibitors induce many side effects, given the essential role of TGF-β signaling in healthy tissues, thereby precluding the treatment of long-lasting diseases such as some cancers, and rare diseases. We thought that we should try a different approach and put the focus on the SMAD proteins, which are the messengers of TGF-β signaling in the cell. The loss-of-function in tumor suppressor genes such as SMAD4 highlights these proteins as potential targets for small-molecule discovery. SMAD4 is present as a single gene, thereby ensuring higher selectivity of the target with respect to R-SMADs and I-SMADs, which, due to gene duplication events in vertebrates, are present as five R-SMADs and two I-SMADs in humans. In addition, a decade of genomics aimed to describe human diseases has revealed other *SMAD4* gene alterations, among them mutations observed in rare diseases such as Myhre syndrome, which is caused by a gain-of-function mechanism that enhances the stability of SMAD4 and alters its roles during tissue regeneration/homeostasis and neural development. Restoring the defects induced by all these mutations poses a therapeutic challenge that requires the identification of starting hit-molecules like those obtained in the EU-OPENSCREEN-Drive and Chem projects, described in the 4.3 section of this thesis. For the screening campaign, we chose the Differential Scanning Fluorimetry (DSF) technique, which allows us to follow the changes in the melting temperature of SMAD4 in the presence of binders. DSF is an accessible, rapid and inexpensive biophysical technique that has found many applications over the years, ranging from the detection of protein folding states to the identification of ligands that bind to the target protein, **(Martin *et al.*, 2013; Gao, Oerlemans and Groves, 2020; Støve *et al.*, 2020)**. One of its major strengths is that the system can perform high-throughput screening using 96- or 384-well plates, facilitating the experimental screening of large libraries of compounds. A second strength is the instrumentation required for the assays. Many laboratories already have (or have access to) real-time polymerase chain reaction (RT-PCR) equipment that allows fluorescence measurements over a controlled temperature range. This eliminates the need for a dedicated instrument. To facilitate the DSF analysis, we have developed the HTSDSF-explorer software together with the group in Bergen **(Martin-Malpartida *et al.*, 2022)**. The software pre-analyzes and displays the T_m and (ΔT_m) results interactively, thereby permitting the user to analyze hundreds of conditions in minutes and select the primary hits. This application also allows the determination of preliminary binding constants, as approximated dissociation constants (K_D=1/binding constant) through a series of subsequent DRAs, facilitating the ranking of validated hits and the advance through the drug discovery challenge. We have also developed a second web application that allows the determination of thermodynamics parameters using the information obtained from DSF assays **(Martin-Malpartida *et al.*, 2024)**. Both applications are available at GitHub (https://github.com/maciaslab).

Using this approach, we have identified 186 novel hit compounds that modify the stability of the WT SMAD4 MH2 domain (either by decreasing or increasing stability) and ranked them based on dose-response assay (DRA) values to guide the next steps of hit-to-lead optimization **(Figure 7)**. Validated hits that bind with good-medium affinity will not be discarded, since they might be derived as new efficient PROTAC molecules or as chemical probes as follow-up projects. These findings are explained in detail in the 4.3 chapter of the Results section.

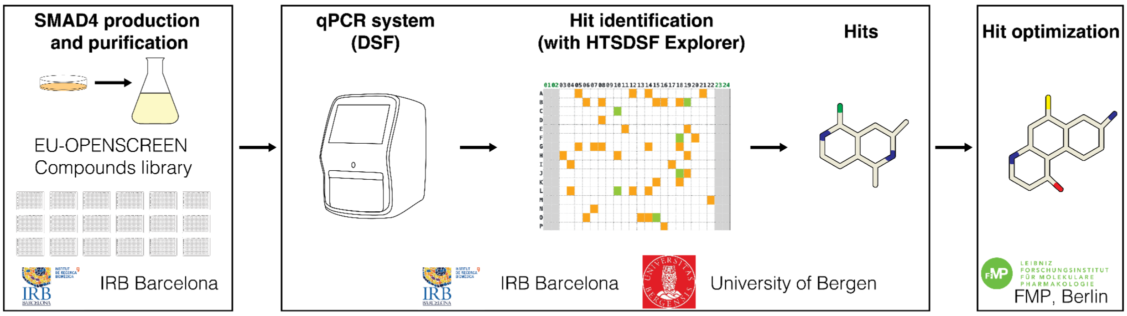

SMAD4 production and purification
EU-OPENSCREEN Compounds library
IRB Barcelona
qPCR system (DSF)
Hit identification (with HTSDSF Explorer)
Hits
IRB Barcelona
University of Bergen
Hit optimization
FMP, Berlin

Figure 7. Schematic representation of the screening strategy thanks to the grants we got to access EU-OPENSCREEN research infrastructures.

1.8 SMAD transcription cofactors and selective regulation of gene expression.

The functional role of SMAD complexes is determined by the expression of many context-dependent transcription partners or cofactors with which they form specific functional transcription complexes. Research carried out during the last two decades has revealed a long list of these cofactors. Examples of modulators of SMAD dependent gene transcription are SKI **(Luo *et al.*, 1999; Tecalco-Cruz *et al.*, 2018)**, SnoN **(Tecalco-Cruz *et al.*, 2018)**, TGIF**(Lo, Wotton and Massagué, 2001; Wotton *et al.*, 2001; Guca *et al.*, 2018)**, P300 **(de Caestecker *et al.*, 2000)**, FOXH1 **(Aragón *et al.*, 2019; Pluta *et al.*, 2022)** or some effectors of EMT gene transcription regulators as ZEB2, OLIG1 **(Motizuki *et al.*, 2013)**, MAN1 **(Pan *et al.*, 2005; Miyazono *et al.*, 2018)** or SNAI1 **(Vincent *et al.*, 2009)** and RREB1 **(Vagne-Descroix *et al.*, 1991; Li *et al.*, 2023)**. The interactions of these cofactors with the SMAD proteins are also structurally described in some cases **(Figure 8)**.

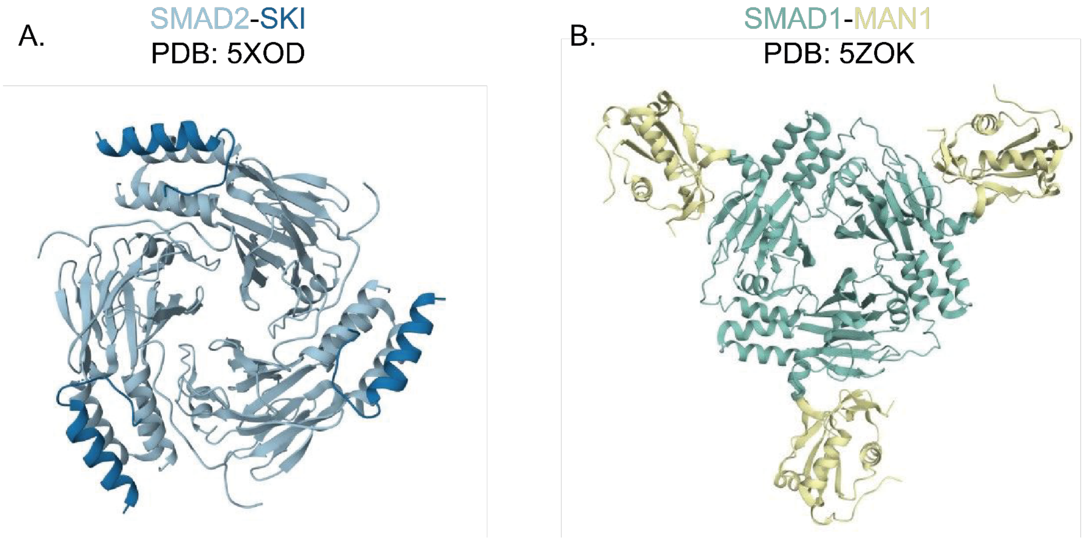


Figure 8. Complexes of R-SMADs bound to cofactors.

A. SMAD2-SKI (PDB:5XOD) and B. SMAD1-MAN (PDB:5ZOK). The crystals contain three units of R-SMADs MH2 domains, each bound to a cofactor. The color code for protein and cofactor is indicated at the top of the panels.

One of the transcription factors attracting interest for its role in the transcriptional regulation of EMT genes and crosstalk with the TGF-β signaling pathway is the Ras-responsive element binding protein 1 (RREB1). RREB1 is a transcription factor that regulates embryo cells' differentiation during gastrulation as well as cell proliferation, transcriptional regulation and DNA damage repair **(Deng *et al.*, 2020)**. RREB1 was originally isolated from thyroid carcinoma cell lines and identified as a transcriptional activator of calcitonin in response to Ras signaling **(Thiagalingam *et al.*, 1996)**. Mammalian RREB1 and the Drosophila orthologue Hindsight regulate epithelial integrity and cell migration **(Yip, Lamka and Lipshitz, 1997; Melani *et al.*, 2008)**. In humans, RREB1 regulates glucose balance whereas the imbalance of RREB1 function plays a role in the development of various cancers and leukemia, as well as in type 2 diabetes, and intervertebral disc degeneration and participates in Zn transport **(Kent, Fox-Talbot and Halushka, 2013; Deng *et al.*, 2020)**. In gastric cancer, RREB1 is highly expressed, and knocking down RREB1 inhibits cell proliferation via increasing p16 expression **(Gao *et al.*, 2021)** . Upon phosphorylation by mitogen-activated protein kinase Ras-MAPK, RREB1 recruits TGFβ-activated SMADs leading to the transcriptional activation of genes that trigger EMT **(Su *et al.*, 2020)**.

RREB1 expression is found in almost all human tissues and cancer cell lines **(The human Protein Atlas, https://www.proteinatlas.org/).** RREB1 is described to act as a transcriptional repressor or activator, depending on the cellular context. At the sequence level, RREB1 is a large multi-Zinc finger (abbreviated as ZF) protein, four times longer than average protein sequences in eukaryotes **(Brocchieri and Karlin, 2005)**. ZFs are the most abundant DNA binding structures found in eukaryotic transcription factors. The RREB1 ZF domains belong to the Cys2-His2 family **(Thiagalingam *et al.*, 1996; Miyake, Szeto and Stumph, 1997; Ming *et al.*, 2013)** and can be grouped into three main clusters, separated by large disordered regions lacking other known structured domains **(Figure 9A-C)**. Areas between ZF clusters can be important for a proper orientation for DNA interaction in cells in the transcriptional complex context, or for PPIs. PXDLS motifs in RREB1 have been proved to contribute to complex formation with the C-terminal binding protein (CtBP) to drive tissue specific transcription in gastrointestinal endocrine cells **(Ray *et al.*, 2014)** .

The presence of alternative splicing processes alters the protein length, with six splicing isoforms reported in human cells, which have different expression distribution among the body **(Nitz *et al.*, 2011)** . The longest RREB1 isoform (isoform **α**) has 1742 amino acids and 16 ZFs in humans. Two isoforms, **ε** and **ζ**, present large deletions, containing either the first N- or the last C- terminal ZFs only.

Due to the presence of numerous isoforms and multiple ZFs, finding the specific interactions between RREB1 and DNA has posed challenges, explaining the numerous motifs and long consensus sites described in the literature for the same protein. The consensus site, known as RAS-responsive element (RRE), is a long and composite motif, where the different positions have distinct degrees of conservation **(Figure 9D) (Thiagalingam *et al.*, 1996)**. In mammals, specific motifs have been identified for the ZF1-5 fragment (GGATGG and GGTGG motifs of the angiotensin gene) and GGTCCT and C4AC2ATC4 sites for the ZF14-15 pair **(Zhang, Zhao and Edenberg, 1999; Date *et al.*, 2004)**. In Drosophila Hindsight, only the C-terminal ZF cluster is described to interact with DNA and the GGT[A/C]C[A/C] and GG[A/C][T/G]GC[T/C] sites **(Ming *et al.*, 2013)** .

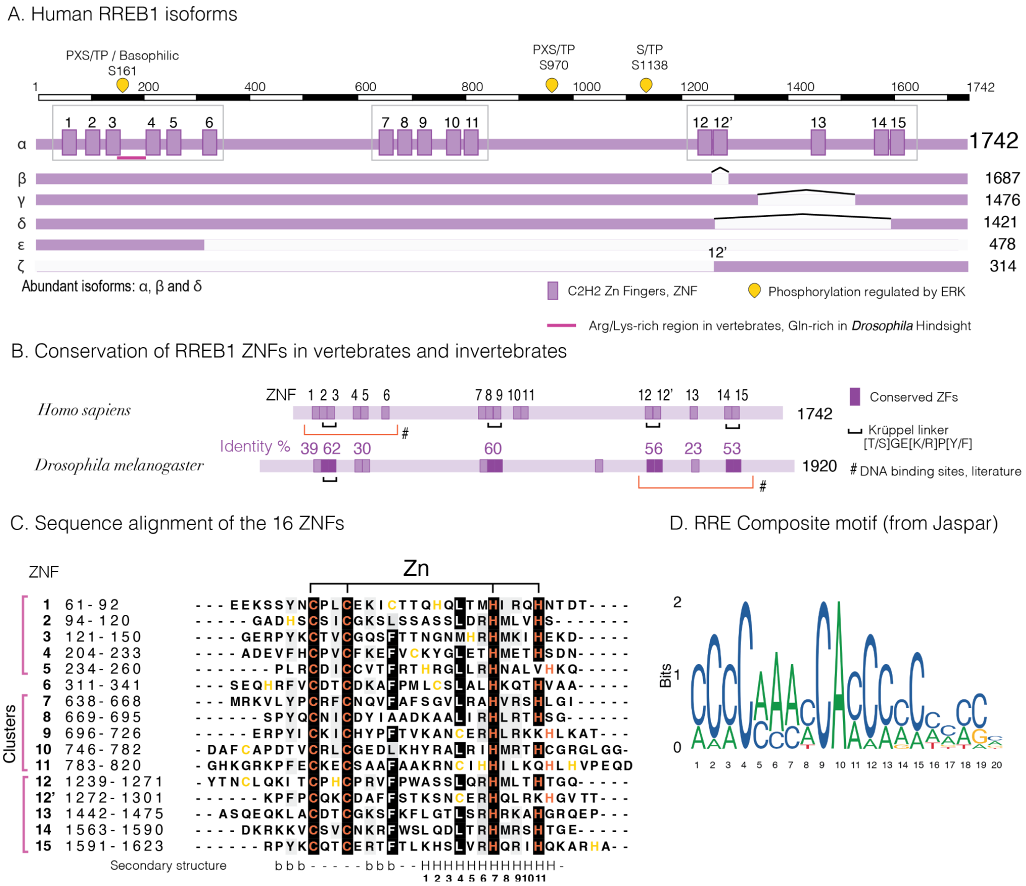
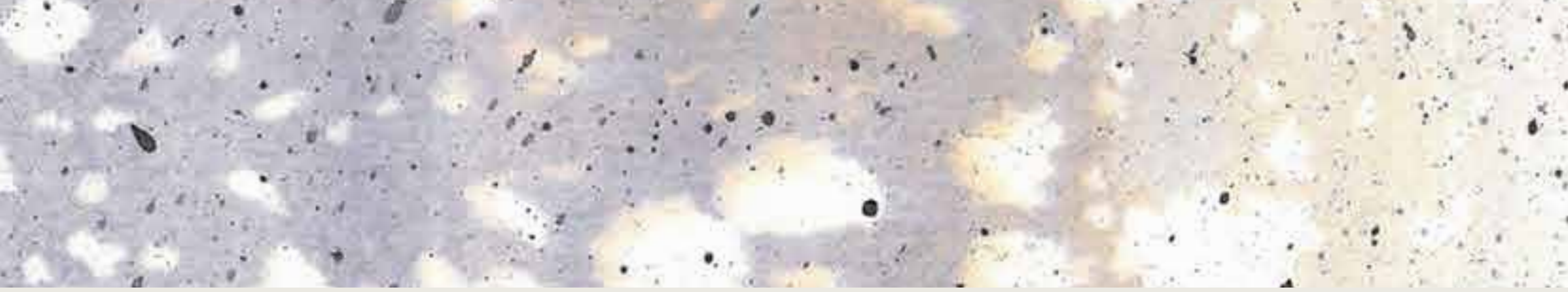

A. Human RREB1 isoforms
PXS/TP / Basophilic S161
PXS/TP S970
S/TP S1138
1
200
400
600
800
1000
1200
1400
1600
1742
1 2 3 4 5 6
7 8 9 10 11
12 12' 13 14 15
α
β
γ
δ
ε
ζ
1742
1687
1476
1421
478
314
12'
Abundant isoforms: α, β and δ
C2H2 Zn Fingers, ZNF
Phosphorylation regulated by ERK
Arg/Lys-rich region in vertebrates, Gln-rich in *Drosophila* Hindsight
B. Conservation of RREB1 ZNFs in vertebrates and invertebrates
ZNF 1 2 3 4 5 6 7 8 9 10 11 12 12' 13 14 15
Homo sapiens
1742
Identity % 39 62 30 60 56 23 53
Drosophila melanogaster
1920
#
#
Conserved ZFs
Krüppel linker [T/S]GE[K/R]P[Y/F]
# DNA binding sites, literature
C. Sequence alignment of the 16 ZNFs
ZNF
Zn
Clusters
1 61-92 ---EEKSSYNCPLCEKICTTQHQLTMHIRQHNTDT----
2 94-120 -----GADHSCSICGKSLSSASSLDRHMLVHS-------
3 121-150 ----GERPYKCTVCGQSFTTNGNMHRHMKIHEKD-----
4 204-233 ----ADEVFHCPVCFKEFVCKYGLETHMETHSDN-----
5 234-260 -------PLRCDICCVTFRTHRGLLRHNALVHKQ-----
6 311-341 ---SEQHRFVCDTCDKAFPMLCSLALHKQTHVAA-----
7 638-668 ---MRKVLYPCRFCNQVFAFSGVLRAHVRSHLGI-----
8 669-695 ------SPYQCNICDYIAADKAALIRHLRTHSG------
9 696-726 -----ERPYICKICHYPFTVKANCERHLRKKHLKAT---
10 746-782 -DAFCAPDTVCRLCGEDLKHYRALRIHMRTHCGRGLGG-
11 783-820 -GHKGRKPFECKECSAAFAAKRNCIHHILKQHLHVPEQD
12 1239-1271 -YTNCLQKITCPHCPRVFPWASSLQRHMLTHTGQ-----
12' 1272-1301 ------KPFPCQKCDAFFSTKSNCERHQLRKHGVTT---
13 1442-1475 --ASQEQKLACDTCGKSFKFLGTLSRHRKAHGRQEP---
14 1563-1590 ----DKRKKVCSVCNKRFWSLQDLTRHMRSHTGE-----
15 1591-1623 ------RPYKCQTCERTFTLKHSLVRHQRIHQKARHA--
Secondary structure bbb-----bbb--HHHHHHHHHHH-
1 2 3 4 5 6 7 8 9 10 11
D. RRE Composite motif (from Jaspar)
Bits
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Figure 9. Domain composition and ZF distribution of human RREB1.

A. RREB1 human isoforms. Experimentally characterized phosphorylation sites are shown in yellow and the characteristic ZF domains are shown as rectangles and numbered. B. Comparison of ZFs groups between Homo Sapiens and Drosophila melanogaster sequences. C. Sequence comparison of the specific ZFs in the largest human isoform with the most conserved positions highlighted. Key cysteine and histidine residues required for zinc coordination are indicated in yellow. D. RREB Composite motif from Jaspar2018 database, profile MA0073.1.

Given the large number of DNA motifs proposed for the protein, we sought to elucidate the specific contacts with DNA to reveal the binding preferences of the C-terminal ZFs. To achieve this aim, we have applied a combination of binding assays and atomic resolution X-ray crystallography. These results are described in the 4.4 chapter of the Results section.



HYPOTHESIS AND AIMS

2.HYPOTHESIS AND AIMS

The ultimate goal of this work is to obtain new knowledge and illustrate key steps in the transforming growth factor β (TGF-β) pathway using molecular, biophysical, structural and chemical biology techniques. The innovative goal is to advance in the development of therapies for cancer and rare diseases, by identifying vulnerable sites in the involved proteins, notably SMAD4, and identifying modifying compounds **(Figure 10)**.

AIMS:

1. Studying the quaternary structures of SMAD proteins and how these structures are affected by a few selected disease associated variants in SMAD4.
2. Using high throughput screening (HTS) of libraries of compounds, identify molecules that interact with SMAD4, including a set of FDA/EMA approved drugs in the context of a drug repurposing screening campaign for cancer, fibrosis and rare diseases.
3. To elucidate the specific contacts with DNA to reveal the binding preferences of the RREB1 C-terminal ZFs, using a combination of binding assays and atomic resolution techniques.

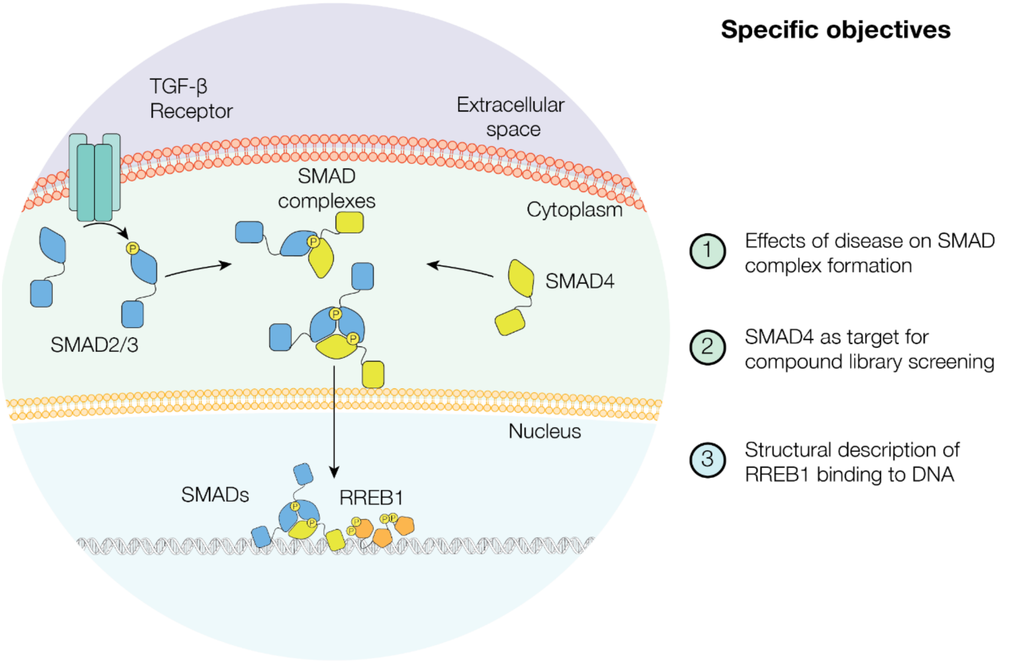
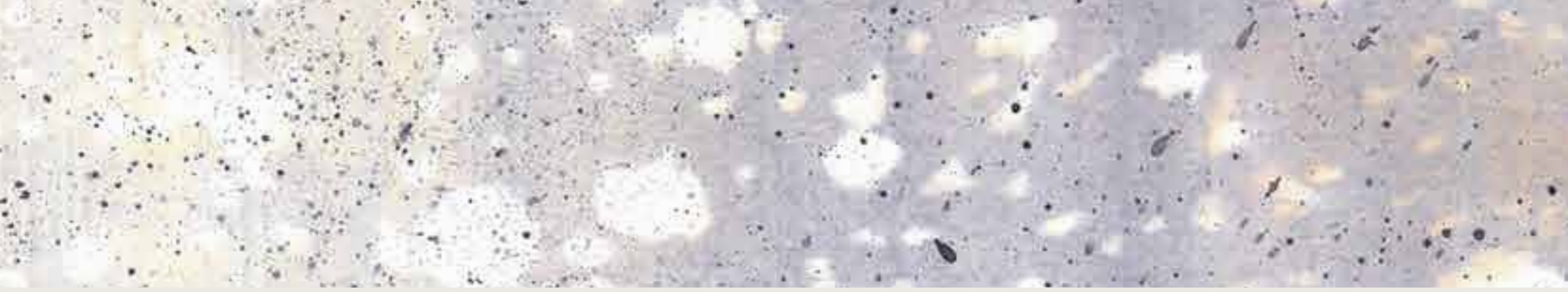
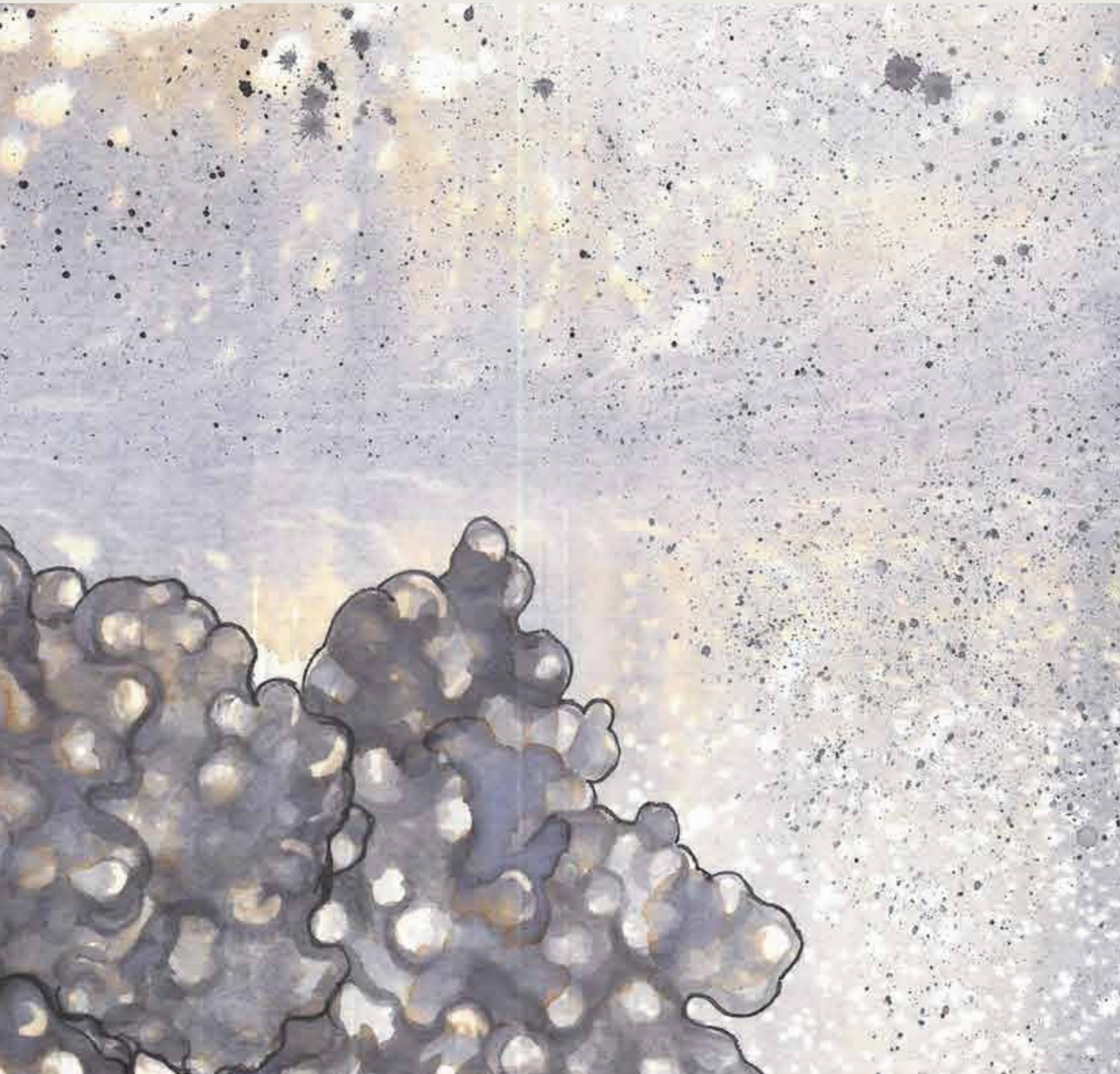

Specific objectives
TGF-β Receptor
Extracellular space
SMAD complexes
Cytoplasm
SMAD4
SMAD2/3
Nucleus
SMADs
RREB1
P
1 Effects of disease on SMAD complex formation
2 SMAD4 as target for compound library screening
3 Structural description of RREB1 binding to DNA

Figure 10. SMAD signaling and interactions: Snail1 and RREB1.



METHODOLOGY



3. METHODOLOGY

3.1 Protein cloning, expression, and purification

SMAD4 SADMH2 272-552 (WT and mutants) were cloned in pETM11, SMAD4MH2 314-552 (WT and mutants) in pOPINS and FL SMAD4 1-550 (WT and mutants) in pCoofy34. SMAD1 MH2 (WT) and SMAD1 MH2EEE (S462E, S463E, S465E), SMAD2 MH2 231-467 (WT) and SMAD2 MH2 EEE (S464E, S465E, S467E) and SMAD3 MH2 189-425 (WT and SMAD3 MH2-DVD, S423D, S425D) were cloned in pOPINF vector **(Table 2).** Define fusion partners, SUMO. Pre-digested pOPIN plasmids were supplied by the Protein Expression and Purification platform at the IRB Barcelona.

All proteins were grown in 2 L Erlenmeyer at 37 ºC, in a combination of LB-TB broth media (20 g LB, 20 g TB, 5 mL glycerol) with 0.05 mg/mL of Kanamycin or Ampicillin. At OD≈0.8, expression was induced with 0.5 mM IPTG and overnight incubation at 20ºC, 200 rpm. Cultures were centrifuged at 3500 g, at 4 ºC. Pellets were resuspended with 40 mM Tris pH 8.0, 40 mM Imidazole, 400 mM NaCl, 0.5% Tween 20 and 1 mM TCEP in all the cases, except for FL SMAD4 constructs, which were resuspended with 100 mM Tris pH 8.0, 150 mM NaCl, 5% glycerol and 1 mM TCEP. Resuspended pellets were incubated with 23 µg/ml Lysozyme and 5 µg/ml DNaseI and the cells were lysed using a pre-cooled Avestin Emulsiflex C3 cell disrupting system. After lysis, the solution was centrifuged for 15’ at 45000-50000 g. Overexpressed proteins were mostly located in the supernatant fraction.

Table 2. SMAD protein constructs used in this work

Construct	Plasmid	Antibiotic	Cleavage	Tags
SMAD4 SADMH2 272-552	pETM11	Kanamycin	TEV	His-Tag
SMAD4 MH2 314-552	pOPINS	Kanamycin	SUMO protease	His-Tag SUMO
FL SMAD4 1-550	pCoofy34	Kanamycin	-	Strep-Tag
SMAD1 MH2 259-465	pOPINF	Ampicillin	3C	His-Tag
SMAD2 MH2 231-467	pOPINF	Ampicillin	3C	His-Tag
SMAD3 MH2 189-425	pOPINF	Ampicillin	3C	His-Tag

His-Tag proteins were purified using a prepacked 5mL His-trap HP column (GE), at 4 ºC in a Bio-Rad NGC chromatography system running at 5mL/min. Proteins were eluted using a buffer with a 40 mM Tris pH 8.0, 400 mM NaCl, 0.5% Tween 20 and 1 mM TCEP and a gradient of Imidazole, 2-400 mM. In the case of the SAD-MH2 protein, we added a 10% buffer step in the middle of the gradient to separate undigested protein from the auto pre-digested one. For SEC-MALS assays, we further purified the SUMO-SMAD4 MH2 using size exclusion chromatography (SEC) using a Superdex75 HiLoad 16/600 column on an ÄKTA purifier FPLC system at 1.5ml/min. Each construction was digested with its corresponding protease **(Table 1)** as follows**.** The His-SUMO tag was cleaved with SUMO at 4 ºC along with buffer exchange, to remove the excess of salt and imidazole. Final buffer was a 20 mM Tris pH 8.0, 100 mM NaCl, 1 mM TCEP. Once the tag was cleaved, the digested His-SUMO tag was separated using a second HisTrap purification step. A final SEC step was performed as above.

SMAD4 MH2, SUMO-SMAD4 MH2, SMAD4 SADMH2 and R-SMAD MH2 domains were purified in 20 mM Tris pH 7.5, 100 mM Tris, 2 mM TCEP and concentrated to 15-20 mg/ml, 18-25 mg/mL, 4 mg/mL and 3-4 mg/mL respectively. Protein purity was confirmed by SDS-PAGE, frozen in liquid nitrogen, and stored at -80 ºC.

Regarding Strep-tag FL-SMAD4, cultures were lysed at 4 ºC in the presence of sigmafast protease inhibitor cocktail (Thermo). Constructs were purified using a Strep-trap column and 150 mM Tris, 200 mM NaCl, 5% glycerol, 1 mM TCEP buffer (washing step) and eluted using 150 mM TRIS, 200 mM NaCl, 5% glycerol, 1 mM TCEP, 2.5 mM desthiobiotin. FL-SMAD4 was purified through SEC as described above and using a 100 mM Tris pH 8.0, 500 mM NaCl, 5%glycerol and 1 mM TCEP as a running buffer to remove the digested tags. For SAXS experiments, we concentrated the samples to 20, 40 and 80 µM.

RREB1 ZF 14-15 pair (residues 1506-1561) was cloned in a pOPINF vector. DE3 *E. coli* strains were grown in LB and induced at OD=0.6-0.8 with 0.5 mM IPTG, followed by O/N incubation at 20 ºC. Lysis was performed using a pre-cooled Avestin Emulsiflex C3 system, lysates were centrifuged, and the supernatants were purified using prepacked 5 mL His-trap HP columns and SEC. Final purification buffer was 20 mM Tris pH 7.5, 100 mM NaCl and 2 mM TCEP.

3.2 Dynamic light scattering (DLS)

Dynamic light scattering (DLS) allows the determination of the hydrodynamic radius of the particles in a solution. Hydrodynamic radius provides information about protein size, and molecular weight, and it was used to measure sample quality and homogeneity. In

DLS, a sample in solution is irradiated with a light source and specific monochromatic weave length. The intensity of the light scattered by the molecules in solution is then recorded and processed. In solution, the constant random movement of molecules (Brownian motion) produces a fluctuation in the scatter intensity. These fluctuations can then be transformed into information about the size of the biomolecules in solution **(Stetefeld, McKenna and Patel, 2016)**.

In this work a Wyatt DynaPro PlateReader II at the ProLinC Facility, University of Linkóping, Linköping, Sweden, and a Nanotemper Prometheus Panta at the SPC facility, EMBL Hamburg, Hamburg, Germany, were used to check sample quality and aggregation.

3.3 Differential scanning fluorimetry (DSF) and nanoDSF

Differential scanning fluorimetry (DSF) is a technique that monitors protein thermal unfolding, which is dependent on the protein sequence and the experimental conditions, such as buffer, additives or small molecules. DSF allows the user to follow the denaturation of a certain protein as a function of temperature monitoring the fluorescence signal of a hydrophobic fluorescent dye that binds to the protein as it unfolds. From this information the melting temperature (T_m) of the protein, which is associated with its stability, can be calculated. Changes in T_m based on e.g. ligand binding can be determined **(Figure 7)**.

Through collaboration with EU-OPENSCREEN and in the lab of Prof. Aurora Martinez, at the University of Bergen, we screened a total of 100037 compounds of the EU-OPENSCREEN library (divided in the Pilot Library and the Diversity Library) and a small library of approved compounds (Prestwick Chemical Library) using DSF **(Støve *et al.*, 2020)**.

DSF has also been used to evaluate changes in the thermal stability driven by mutations in protein sequence, evaluate protein-protein interactions, protein-peptide interactions and protein-small molecule interactions.

Thermal stability assays were performed using QuantStudio6Flex 384-well plates qPCR, Roche LightCycler 480 II or BIORAD CFX384. In each case, different melting curves were used:

- QuantStudio6Flex: Temperature stabilization at 25 ºC during 1 min. This step is followed by a melting curve from 25 to 99 ºC with an increment of 0.3 ºC each 8s. Final step of 15 s at 99 ºC.

- Roche LightCycler 480 II: Melting curve from 20 to 99 ºC with 0.04 ºC/s temperature increase with 4 acquisitions at each temperature when using the equipment in the University of Bergen or a melting curve from 25 to 85 ºC.
- BIORAD CFX384: Melting curve from 20 ºC to 99 ºC with a 0.5 ºC temperature increase at each step.

Assays were performed using 0.5 mg/ml of SMAD4 272-552 construct and 5x Sypro Orange in a final volume of 10 µL or 25 µL respectively of each qPCR system. Samples were prepared in 20 mM Tris pH 7.5, 100 mM NaCl, 1 mM TCEP, 5x Sypro Orange stock was prepared from a 5000x stock (Merck) and diluted with a protein buffer. Results were analyzed using HTSDSF Explorer **(Martin-Malpartida *et al.*, 2022)** and Microsoft Excel.

NanoDSF relies on the intrinsic tryptophan fluorescence signal. In a nanoDSF experiment, tryptophans and other aromatic residues are excited by UV light at 280 nm and its fluorescence emission recorded. Aromatic residues emit at 330 nm when protected from an aqueous environment, but have an emission fluorescence spectral shift to 350 nm when exposed. Upon unfolding then, the enriched hydrophobic core of the protein (containing aromatic residues) will change its chemical environment, with more water molecules, increasing the 350 nm fluorescence. Following the changes in the 350 nm/330 nm fluorescence ratio, one can subtract the inflection point values or T_m values of the unfolding curve. At the same time, a similar approach can be done using the raw 350 nm or 330 nm fluorescence. Some of the advantages that nanoDSF can offer with respect to DSF is to avoid using Sypro Orange dye, thus diminishing the experimental artifacts that may derive from the potential non-specific interaction of this molecule with hydrophobic patches of the target protein, resulting in erroneous T_m values.

In this work, condition optimization was performed in a Prometheus NT.48 (Nanotemper) from the ProLinC Facility, University of Linköping and final measurement performed in Prometheus Panta (Nanotemper) from the SPC facility, EMBL Hamburg.

20 µL of samples were prepared at the desired concentrations in low binding 200 µL tubes. After samples were prepared and incubated, they were centrifuged to remove air bubbles, and loaded inside the Prometheus NT.48 Series nanoDSF Grade High Sensitivity Capillaries. Prometheus capillaries have to be completely filled, being especially careful with air bubbles. In the case that an air bubble could not be removed, they were displaced to the sides of the capillary.

Final protein unfolding experiments in Prometheus systems were performed with a temperature slope of 1 ºC/min from 25 to 80 ºC. The used excitation power was automatically selected in the “Discovery Scan mode”. Excitation Power parameter is selected depending on the amount of fluorescent signal determined by the number of tryptophans and, by far less extent, phenylalanines and tyrosines.

3.4 Mass photometry (MP)

Mass Photometry (MP) also known as interferometric scattering mass spectrometry (iSCAMS), is a recently developed technology which allows the measurement of macromolecules molecular weight using the light scattered and reflected in a glass surface. Purity, aggregation, stoichiometry and oligomerization are some of the properties which can be studied through this method. More advanced applications, involving protein oligomerization, interaction affinity and binding rates, are also possible. Different ranges of molecules from proteins and nucleic acids to whole viruses are measurable.

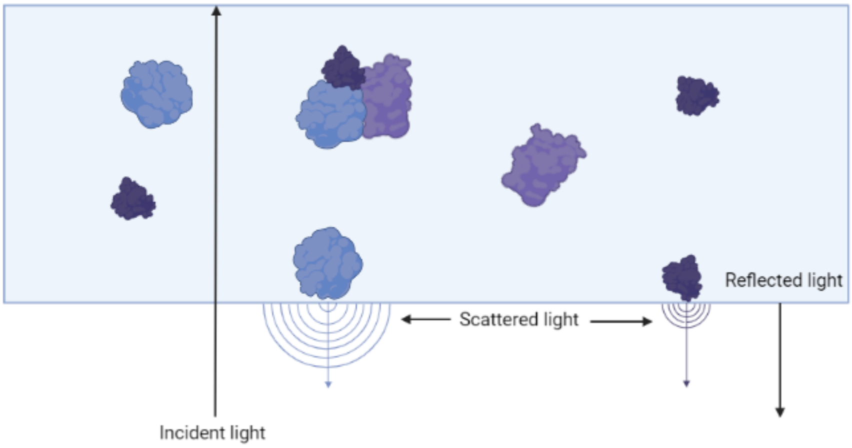


Figure 11. Diagram of a mass photometry experiment.

Reflected and scattered light are measured to subtract the interferometric contrast. Picture adapted from **(Young *et al.*, 2018)**.

In an MP experiment, a sample containing biomolecules, or their complexes, is irradiated with light. As a result, the component in solution will produce scattered light, while light will be reflected by the measurement surface **(Figure 11)**. The interference of them is detected and processed to obtain the interferometric contrast of the particles in solution, which is directly related to their molecular weight. This way we can obtain the molecular weight and also the mole ratio **(Young *et al.*, 2018; Sonn-Segev *et al.*, 2020)**. Limitations arise from mixtures of different size particles that will have different diffusion coefficients. In fact, in an MP experiment, large molecules will always be easily detected and be more enriched, since they have a lower diffusion coefficient **(Sonn-Segev *et al.*, 2020)**.

The MP methodology was developed by Philipp Kukura's lab in University of Oxford, which led to the Refeyn Ltd company that commercialized these systems. Measurable molecular masses will be limited by the experimental set-up of the mass-photometer and differ in resolution.

- Refeyn OneMP: Molecules from 40 kDa to 5 MDa

- Refeyn TwoMP: Molecules from 30 kDa to 5 MDa

- SamuxMP: High molecular weight particles like virus capsids.

The mass photometer set-up is composed of a system of lenses, a polarized beam splitter, a camera and a 445 nm laser **(Figure 12)**.

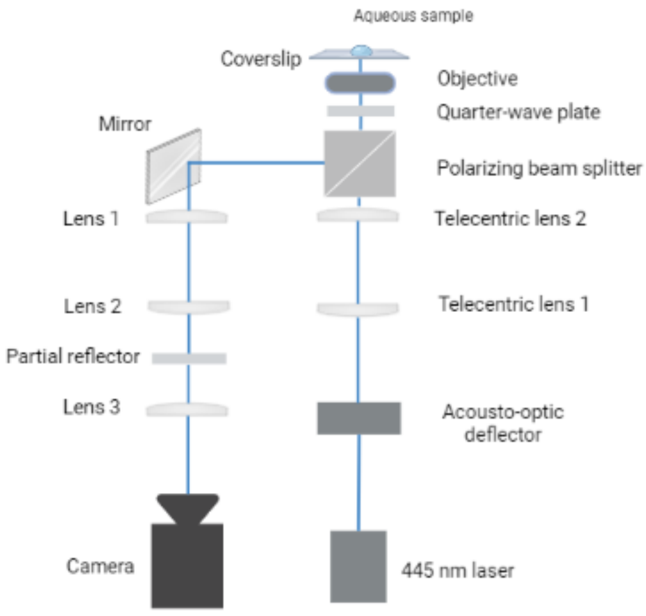


Figure 12. Typical set-up in a mass-photometer.
Adapted from Refeyn website and patent US10816784B1.

The method relies on a calibration curve with macromolecules of known species and molecular masses. In the case of proteins, it is assumed that each amino acid type produces a similar scatter, allowing for the comparison of proteins. Reference proteins are listed below:

- β-Amylase (A8781, Merck). It forms three species of 56kDa, 112 kDa and 224 kDa.
- Bovine Serum Albumin (BSA) (23209, Merck). Monomer of 66kDa and dimer 132 kDa.
- γ-Globulin (9007-83-4, Merck). Monomer 150 kDa and dimer 300 kDa species.
- Thyroglobulin: a 670kDa monomer.
- NativeMark™ unstained protein standard (NM), catalog number LC0725, Life Technologies. We used this standard in our experiments.

In MP, an acceptable error is in the range of ±5% of the theoretical molecular weight of the system. It is advisable to acquire several replicates of the same sample to minimize the experimental errors. Other sources of errors include a deviation in the calibration measurements or due to bad fitting of the raw data.

3.4.1 Pipeline for Mass photometry experiments

In our MP experiments, we were following the next pipeline optimized by SPC Facility, EMBL Hamburg:

1. Commercial coverslips cleaning. Sonicator baths in water and isopropanol are recommended to remove particles from the glass surface (15 min in mQ water, 15 min in isopropanol and 15 min in mQ water).
2. Mount and placement of cover slips and Refeyn sample well cassette. Both elements are mounted using a special provided mold and positioned on the top of the objective/tray of the MP system. Before that, it is needed to add on the top of the objective a drop of oil suitable for optical systems.
3. The first step for a mass-photometry measurement is the positioning of the optic system under the well with the sample. Positioning the optics a little away from the well center improves the measurements a bit.
4. Application of an 18 µL buffer drop and adjustment of sharpness parameter (higher sharpness increases image resolution). Once the correct position is found, the focus is selected and locked manually. Ideal positions can be found automatically by the software. Focus locking has to be performed in a position in which native view and ratiometric view show no particles in presence of buffer.
5. Measurement of buffer control. This measurement has to show a low number of counts. 80-150 counts at signals greater than 0.07 are the accepted maximum.
6. Calibration curve. Addition of 2.5 µL of NM protein standard and drop homogenization using a P20 micropipette. Measurement must be done

immediately to obtain a valid calibration curve (NM1:66 kDa, NM2: 146 kDa, NM3:480 kDa, NM4:1048 kDa).

7. Sample measurement. 1 µL of 50 to 100 nM samples are loaded into the buffer drop and homogenized with a P20 micropipette. Final protein concentrations in our assays were 5.26 nM (SMAD3) and 2.63 nM (SMAD4).
8. Cassette change. Every two or three cassette changes, more oil should be added at the top of the optics. Previous cleaning with folded MC-50E Lens Tissues (ThorLabs) soaked in isopropanol may be needed. Cassettes, but not cover slips, can be reused. Cassettes will be stored in a 50 mL falcon filled with isopropanol and cleaned afterward with the same procedure as for the cover slips.
9. Cleaning of the optic system. Use MC-50E Lens Tissues soaked in isopropanol (to prevent scratches) to clean the optic system. Perform cleaning after every set of experiments.

In our experiments we used a 20 mM Tris pH 7.5, 100 mM NaCl buffer filtered with Steriflip® Filter Units (Merck).

3.5 Surface Plasmon Resonance (SPR)

Surface Plasmon Resonance (SPR) assay is a biophysics method that allows detecting a binding reaction of two or more macromolecules in real-time as well as measuring parameters such as dissociation constant (K_D), binding rates (k_{on} and k_{off}) and thermodynamic parameters.

Surface plasmon resonance occurs when photons strike at a thin metal film (Chip) at an angle that depends on the refractive index of the material near the metal surface. This material can vary depending on the plate used for the assay. When this happens, the electrons in the metal film are excited and move. These electron motions are called plasmon, and they propagate through the metal film. We can attach a biomolecule, such as a protein, to the opposite side of the metal plate from the incident light. In this situation, a change in the state of the protein, such as when it binds a ligand, would change the refractive index of the material and interfere with the formation of plasmon, changing the amount of reflected light. This difference can be measured and correlated with the binding or kinetic properties of the biomolecular system under study **(Figure 13A)**.

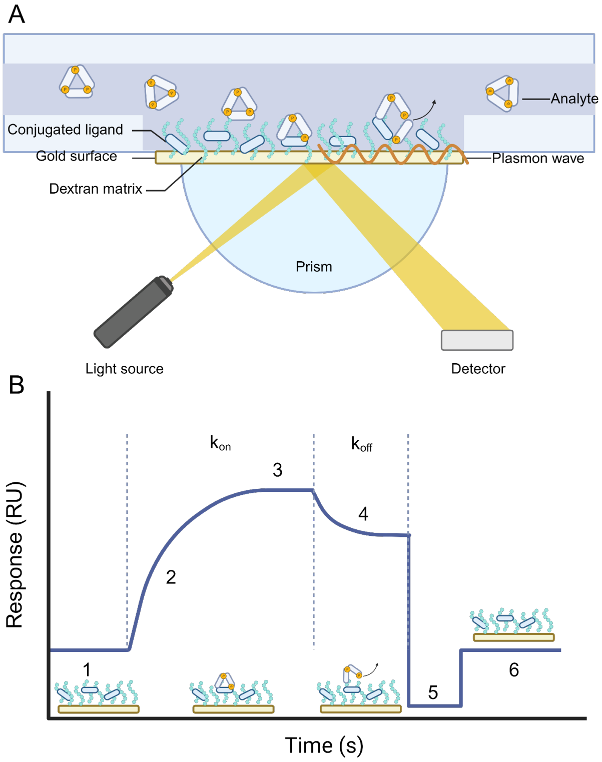


Figure 13. Diagram of SPR detection system.

A. Experimental set up for SPR experiments involving SMAD proteins is shown, where monomeric SMAD4 is immobilized on the chip surface, while activated SMAD3 is used as analyte. **B.** Sample sensorgram. Phases: 1- Baseline, 2- Association, 3- Steady-state, 4- Dissociation, 5- Regeneration, 6- Baseline. RU represents resonance units.

In SPR systems, a macromolecule is immobilized in a chip surface (the ligand) and then titrated with increasing concentrations of an interaction partner or a possible interaction partner (the analyte) thanks to a very precise microfluidic system. The immobilization of the biomolecule can be either covalent or non-covalent.

After immobilization, a microfluidic system is used to inject the analyte. Binding events, such as association and dissociation, will produce changes in the amount of reflected light. The plot of this signal versus time is called a sensorgram and shows these events **(Figure 13B)**. Fitting each phase of the sensorgram curves to a different model would allow us to extract the kinetic and thermodynamic parameters of the binding events. Finally, during the regeneration phase, a slightly more aggressive buffer is used to ensure that all the analyte is removed, while the bound protein is kept in the chip and can now be reused for another experiment.

Multi-cycle kinetics experiments were performed with immobilized SMAD4 SAD-MH2 variants titrated with R-SMADs (SMAD3 MH2 phopsphomimetic mutant and a SMAD2 MH2 construct). Experiments were performed in a Biacore T200 (Cytiva) from *Centres Científics i Tecnològics de la UB* (CCiTUB) immobilizing SMAD4 constructs through amine-coupling on a CM5 S Series sensor chip (Cytiva). The recommended immobilization levels for PPIs measurements are 100 RU. In our case, 100 RU gave low signal levels, and after optimization, we used values of 439 RU, 505 RU and 443 RU for WT, I500V and R496C variants. Analyte samples were dialyzed in running buffer O/N at 4 ºC. Running buffer was 20 mM HEPES pH 7.4, 150 mM NaCl, 0.05% Tween 20 and 2 mM TCEP, which was previously filtered through a 0.22 µm filter. We used 5 different concentrations for each analyte, ranging from 1.56 to 25 µM, and we did a global analysis for the five concentrations.

Association and dissociation phase times were 120 s and 180 s, respectively. Flow rate was 50 µL/min (using the low sample consumption mode of the system) at 25 °C. Regeneration phase was performed with a 1.5 M NaCl, 0.5% Tween 20 solution at 30 µL/min flow rate for 30 s.

For data analysis, the Biacore T200 Evaluation Software from Cytiva was used. Due to the complexity of the kinetics, in which the R-SMADS are in an equilibrium between monomer, dimer and trimer, we were not able to fit the data for the association phase. For this reason, we set to analyze only the dissociation data between second 120 and 220 and fitted into a 1:1 dissociation equation with the equation:

$$R = R_0\, e^{-koff(t \cdot t_0)} + Offset \tag{3.1}$$

where R is response (RU), R_0 is the response at dissociation time 0 or at 120 s of the SPR run (RU), k_{off} is the dissociation rate (s^{-1}), t_0 is the injection stop time (s) and the offset is the residual response above baseline after complete dissociation.

The dissociation rate value can be transformed to complex half lifetime ($t_{1/2}$) with the following equation:

$$t_{1/2} = \frac{ln2}{k_{off}} \tag{3.2}$$

3.6 Electrophoretic shift assay (EMSA)

We use Electrophoretic shift assay (EMSA) to detect binding of two or more molecules in a native gel **(Hellman and Fried 2007)**. The sample migrates according to its molecular weight and charge, in contrast to what we observed in SDS-PAGE, where charges are homogenized by SDS detergent and samples migrate according to their size. Protein-DNA complexes are not in equilibrium after sample loading, since species separate while they migrate through the gel. These phenomena traduce into smears in the electrophoretic gel, which correspond to dissociation of low affinity complexes.

In our experiments, we performed the analysis of protein-dsDNA binding following the signal of a Cy5-labeled DNA (λ_{ex}: 649 nm, λ_{em}: 667 nm). Control dsDNA oligos and dsDNA-protein complexes migrate from the negative to the positive pole of the electrophoretic system, resolved according to the specific acrylamide gel used and buffer. With this type of labeling, only DNA (free or bound to the protein) is visible in the experiment.

To prepare duplex DNAs, complementary strands were annealed using HPLC purified DNAs purchased from Condalab. DNAs were mixed at equimolar concentrations (3 mM) in 20 mM Tris pH 7.0 and 10 mM NaCl, heated at 90 °C for 3 min and cooled down to room temperature for 2 h.

Protein and DNA were incubated for 15 min at 4 °C in 10 µL of reaction volume using a protein buffer to dilute the samples. After incubation, we add 10 µL of loading buffer containing orange G and samples are immediately loaded.

We used 1.5 mm thick gels prepared with Tris-Glycine buffer. For RREB1 we used 6% acrylamide (40% 19:1 acrylamide/bis-acrylamide solution from BioRad). After loading samples, gels were run at 110V and 4 ºC in Tris-Glycine buffer. The DNA is kept at 7.5 nM concentration and the protein is added at increasing concentrations in ranges that allow us to detect the complex formation. Gels were analyzed using a Typhoon imager (GE Healthcare) and a Cy5/Red fluorescent filter.

For affinity quantification, bands corresponding to dsDNA and dsDNA-protein complexes were quantified by ImageJ and were used to calculate the fraction bound as follows:

$$f_b = \frac{[dsDNA]_{bound}}{[dsDNA]_{bound+unbound}} \tag{3.3}$$

Fraction bound dependent on protein concentration was fitted with a non-linear regression fit in a *One site - Specific binding* equation in Graph Pad Prism

$$f_b = \frac{B_{max} \cdot P}{(K_D + P)} \tag{3.4}$$

where f_b is fraction bound, B_{max} is the maximum observed fraction bound and P is protein concentration.

3.7 Fluorescence Spectral Shift

A fluorophore, or fluorescent dye, is a fluorescent chemical compound that can emit light upon light excitation. Changes in the chemical environment of a fluorescent dye often produce changes in its emission spectra, **(Figure 14)**. This principle can be used to observe binding reactions between a labeled biomolecule or a fluorescent small molecule with another particle. The output of the assay is a ratio of the intensity of fluorescent light at two different wavelengths. Such changes can be caused by direct interaction of a molecule with the dye, by proximity binding, or by conformational changes that could alter the position of molecules around the fluorophore or the fluorophore itself.

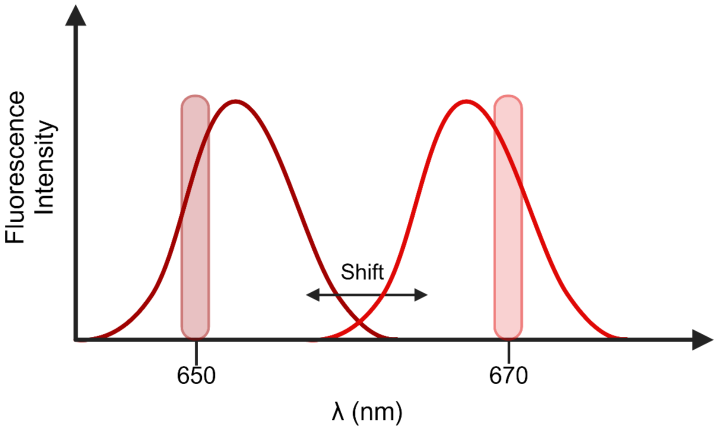


Figure 14. Diagram of the spectral shift change observed in Nanotemper Red.

In this thesis, spectral shift is used to evaluate the binding of a HisTag-SMAD4 272-552 construct, fluorescent labeled with a specific HisTag labeling kit (His-Tag Labeling Kit RED-tris-NTA 2nd Generation, Nanotemper), with small-molecules ligands.

The protocol for labeling includes these steps:

1. Protein buffer. We used HBS-T (50 mM HEPES pH 7.4, 150 mM NaCl, 0,05% Tween. PBS-T is also recommended. Other buffers like Tris or the presence of DTT or TCEP interfere with the labeling.
2. Red-tris-NTA dye affinity should be performed to optimize the amount of reagent (expensive). The titration is performed with a constant concentration of dye and increasing concentration of HisTag-protein.
3. We have found that an efficient dye:protein ratio concentration of 50 nM-100 nM works satisfactorily for a dissociation constant (K_D) equal to or less than 10 nM.
4. After 30 min incubation at room temperature, removal of the dye excess is required and also of protein precipitated or aggregated, by 5 min centrifugation at 15,000 g.

In our assays, the final protein concentration was 25 nM in 20 µL. Plate preparation was done printing the compound with an ECHO liquid handler, followed by buffer addition and 10µL of labeled protein solution. HBS-T buffer was used with 2% DMSO.

For single-dose assays, we used compound solutions at 50 µM and were run as replicates at 25 ºC. Hits were further validated through dose-response assays. Positive compounds were run as replicates and a control was performed with a HisTag peptide. The readout of the experiments was done in a Dianthus Pico from Nanotemper of the IRB Drug Screening platform, which reads fluorescence in two different wavelengths, 650nm and 670nm. Analysis of the data was performed using both DI.Control software and DI.ScreeningAnalysis software from the same company.

The dissociation constant in DRA assays was calculated through the following equation equation **(Langer *et al.*, 2022)** :

$$f_{bound} = \frac{([L]+[T]+K_D-\sqrt{([L]+[T]+K_D)^2-4[L][T]}}{2[T]} \tag{3.5}$$

where f_{bound} is the fraction bound and [L] and [T] are the ligand and target protein concentrations, respectively. K_D then is estimated by fitting the equation:

$$R_{total} = R_{unbound} + f_{bound} \cdot (R_{bound} - R_{unbound}) \tag{3.6}$$

where R_{total} is the 670nm/650nm fluorescence ratio measured with Dianthus Pico at a given concentration of ligand, $R_{unbound}$ is the ratio value of the target alone and R_{bound} is the ratio value of the complex.

3.8 Microscale Thermophoresis (MST) and Temperature-Related Intensity Change (TRIC)

Microscale thermophoresis (MST) follows changes in the mobility of a fluorescent molecule upon the irradiation of a sample with a laser beam. The sample is placed in a cylindrical well in multiwell plates specially designed for this technique. Each experiment requires two wells, one with ligand and one without, which are irradiated with a laser incident in the center of the well. This causes the proteins to move by thermophoresis, which is the movement of the molecules by a temperature gradient, from the center to the edges. The size and shape of the protein will cause this to happen at different rates. When the ligand interacts with the labeled protein, its mobility changes as there is a huge increase in the molecular weight, dynamics, conformation and charge distribution. A second phenomenon in this technique is TRIC (Temperature related intensity change), in which the changes to fluorescence are not related to the movement of the molecule, but to the intrinsic property of the fluorophore to change its intensity in function of the temperature.

The output of this assay is either TRIC (Temperature-related intensity change) or MST traces **(Figure 15)**. In TRIC traces, an initial fluorescence ($F_{t=0}$ or F_0) signal is measured corresponding to the target fluorescence. This fluorescence reduces its intensity upon IR-laser activation ($F_{t=1s}$ or F_1) because of a temperature dependent displacement of the molecules in the solution, which recover initial values after laser inactivation. When a ligand is bound, it can increase the relative fluorescence signal during IR laser activation if it causes the target to slow down or decrease it if it increases the target's speed. In the first scenario, there is a significant change in the molecular weight of the target, while in the second scenario, there is a significant change in the molecular dynamics of the target.

We run these experiments at the same time as we acquired the spectral shift datasets, as the Dianthus Pico can acquire both experiments in the same plate serially.

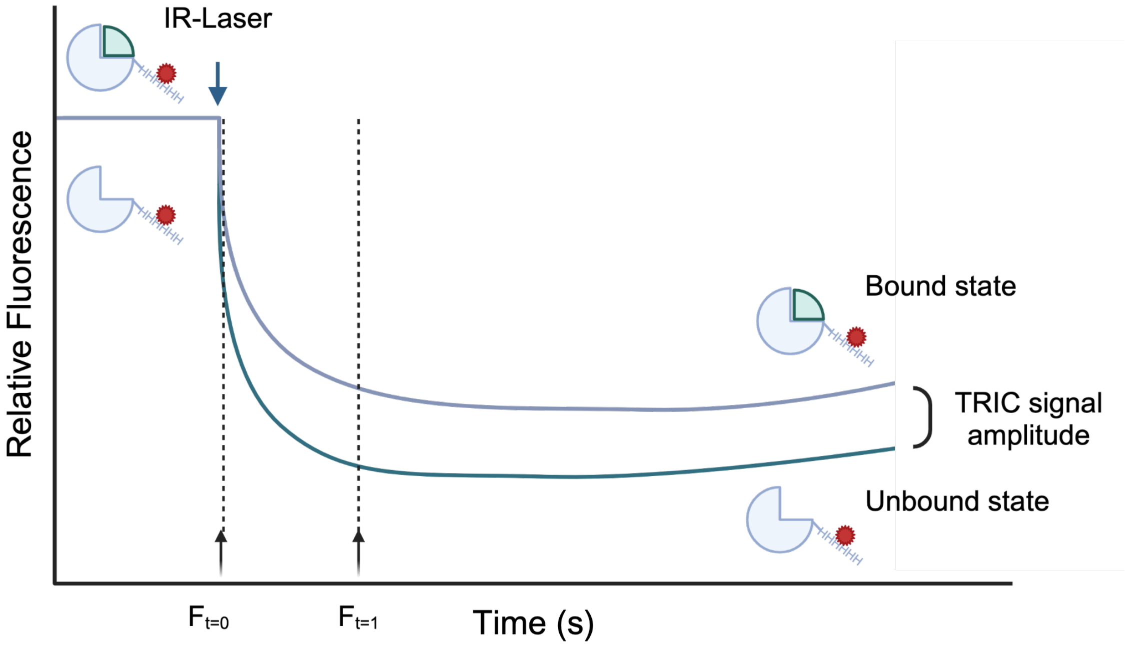


Figure 15. Diagram of TRIC experiment principle.

Upon irradiation with an IR-laser, the labelled molecules are displaced by a gradient of temperature and fluorescent signal is reduced. The speed of signal decay depends on parameters such as molecular weight, shape and overall-charge. Upon binding reaction, some of these parameters may change and increase or decrease the speed at which the fluorescence signal is lost from the center of the well or the capillary where the detector is placed.

In the case of TRIC experiments, the concentration-dependent F_{norm} variations are recorded. F_{norm} is a parameter subtracted by dividing F_1 by F_0, where F_1 is the normalized fluorescence at a given time and F_0 is the normalized fluorescence intensity prior to IR laser activation.

Total normalized fluorescence (F_{norm}) is the sum of bound and unbound normalized fluorescence multiplied by their fractions (f) in the protein populations.

$$F_{norm} = (1 - f_{bound})F_{unbound} + f_{bound}F_{bound} \tag{3.7}$$

K_D values from the titration data can be derived from the fraction of labeled target bound to the ligand. This is done in the same way as in spectral shift experiments, according to the expression:

$$f_{bound} = \frac{([L]+[T]+K_D-\sqrt{([L]+[T]+K_D)^2-4[L][T]}}{2[T]} \tag{3.8}$$

3.9 Complex formation validation trough SEC-MALS

In contrast to the previous techniques, which were used for protein-small molecule interactions, Size-exclusion chromatography coupled with Multi Angle Light Scattering (SEC-MALS) allows the separation of molecules by its molecular weight (depending on the SEC column used) and shape and the calculation of their median molecular weight (MW). This technique is normally used in structural biology to characterize the formation of macromolecular complexes before structural studies. In our case, we used it to characterize the quaternary structure of SMAD proteins. SEC-MALS can measure the molecular weight of a specific molecule or complex thanks to the combination of three different detectors:

- Multi Angle Light Scattering (MALS) detector: Measures the light scattered at multiple angles by analytes in a given elution volume.
- Differential Refractive Index (dRI) Detector: Allows measurement of protein concentration based on changes in the refractive index of the solution caused by the presence of the molecules injected into the chromatographic system at a given elution volume.
- UV 280 nm detector: Enables measurement of protein concentration, which correlates with absorbance at 280 nm.

From the data it is possible to calculate the MW of a certain analyte as follows

$$M = \frac{R(0)}{Kc(\frac{dn}{dc})^2} \tag{3.9}$$

where M is the average molecular weight of the species in a certain elution volume, R(0) is the amount of light scattered extrapolated to angle zero, c is the concentration determined by the UV or dRI detectors, dn/dc is the increment of diffracting index compared with the used buffer and K is a physical constant for the vertical polarized incident light **(Wyatt, 1993)**.

In our case, protein-protein interaction analyses using SEC-MALS was performed to corroborate the complex formation, using a modular HPLC Prominence system from Shimadzu connected to an autosampler SIL-20AC (Shimadzu), a LCD20-ADsp pump (Shimadzu), a SPD-20A UV detector (Shimadzu), a Dawn Heleos II (Wyatt) MALS detector and an Optilab t-Rex RI detector (Wyatt). For the assays, we selected a Superdex 200 10/300 Increase (Cytiva). Buffer was filtered through a 0.22 µm filter before column equilibration steps. Samples were prepared from frozen protein stocks. Proteins were defrosted in ice and centrifuged at 17000 g and 4 ºC during 10 min. After sample preparation, samples were filtered with 0.22 µm membrane filters and stored in the SIL-20AC autosampler. Proteins were combined in 2:1 ratio (SMAD4:SMAD3) to facilitate system saturation even in loss-of-function SMAD4 mutants. 130 µL of injected volume with a 50 µM:25 µM ratio and 190µL of injected volume with a 100 µM:50 µM ratios were used for different runs specifications. Running buffer was 20 mM Tris pH 7.5, 100 mM NaCl, 2 mM TCEP. Results were analyzed using ASTRA (Wyatt) and Microsoft Excel.

3.10 Isothermal titration calorimetry (ITC)

Isothermal titration calorimetry is considered the gold standard to quantify macromolecular interactions and extract accurate affinity and thermodynamic values associated with the binding reaction. The method relies on the measurement of the heat released or absorbed during binding. This can be done in calorimeters that are able to detect changes in the temperature of a solution compared to the one in a reference cell **(Bastos and Velazquez-Campoy, 2021)**.

ITC experiments are performed in a chamber isolated with an adiabatic jacket and two cells regulated by a combination of temperature detectors and heaters. In a typical ITC experiment, a solution is placed in a syringe, and it is progressively injected into the sample cell. The tip of the syringe is also a stirrer and helps to mix its content with that of the cell. Once solution A (syringe) starts interacting with solution B (previously loaded in the sample cell) the differences in temperature between the sample cell and the reference cell is recorded and equilibrated thanks to the action of a feedback heater in the sample cell. Reference cells are kept at a constant temperature.

ITC is based on measurement of heat (q) exchange. The heat released or absorbed is equal to the change in enthalpy (ΔH) when the system is at constant pressure (P). Because enthalpy is directly related with changes in internal energy (ΔE) enthalpy can be expressed as

$$\Delta H = \Delta E + P\Delta V \tag{3.10}$$

$$\Delta E = q + w = q - P\Delta V \tag{3.11}$$

$$\Delta H = q_p - P\Delta V + P\Delta V = q_p \tag{3.12}$$

where ΔV is the change in volume, w is work and q_p is heat at constant pressure.

The heat absorbed or released by the system upon each sample injection is recorded and subsequently, the areas of each peak produced by this heat exchange are subtracted. ΔH, binding constant (K_A) and stoichiometry (n) can be obtained from fitting the areas versus the mole ratio through different equations. For the analysis of the assays shown in this thesis, an *Independent binding model* was applied. The Independent model is suited to explain 1:1 interactions as well as more complex systems as two targets interacting with one ligand molecule or two ligand molecules interacting with one target molecule.

K_D of the system can also be obtained, since it is defined as $1/K_A$. K_A and ΔH permit the determination of other thermodynamic parameters of binding since

$$\Delta G^0 = -RTlnK_A \tag{3.13}$$

And

$$\Delta G^0 = \Delta H - T\Delta S \tag{3.14}$$

where ΔG^0 is the Gibbs Free Energy of binding and ΔS is the binding entropy.

We applied ITC to study SMAD PPIs and RREB1 and DNA interactions. ITC measurements were performed using a nano ITC calorimeter (TA Instruments) at 37 °C and 260 rpm.

In the case of SMAD complexes, 50 µL of 100 µM or 280 µM of SMAD4 314-552 (syringe) and 300 µL of 30 µM or 80 µM SMAD3MH2 DVD (cell) were used respectively for gain-of-function mutants and WT/A406T constructs. The volume of the titrant was divided into a total of 17 injections. Concentrations were re-measured after degassing the samples for possible changes due to evaporation using a NanoDrop system. All proteins were previously dialysed in the same 20 mM Tris pH 7.5, 100 mM NaCl and

2mM TCEP buffer overnight at 4 ºC. The NanoAnalyze software (TA Instruments) was used to analyze the binding isotherms. Baseline controls were acquired with buffer and SMAD4.

RREB1 binding was determined at 25 °C and stirring at 220 rpm as these conditions allowed us to stabilize the system quickly. 50 µL of 115.8 µM GGTCCT DNA was titrated (17 injections) into an 11.7 µM RREB1 ZF14-15 solution (300 µL). Concentrations were determined using a NanoDrop system and their predicted extinction coefficients after degassing the samples. Both DNA and protein samples were dissolved in the same buffer. The NanoAnalyze software (TA Instruments) was used to analyze the binding isotherms. Baseline controls were acquired with buffer and pure DNA solution. Fittings were performed using the independent binding sites model.

3.11 X-ray crystallography

X-ray crystallography has been used in this work to determine the structure of SMAD4 variants, as well as to investigate the interaction of small molecules with variants and WT SMAD4. We have also used this technique to identify the main interaction between RREB1 and DNA.

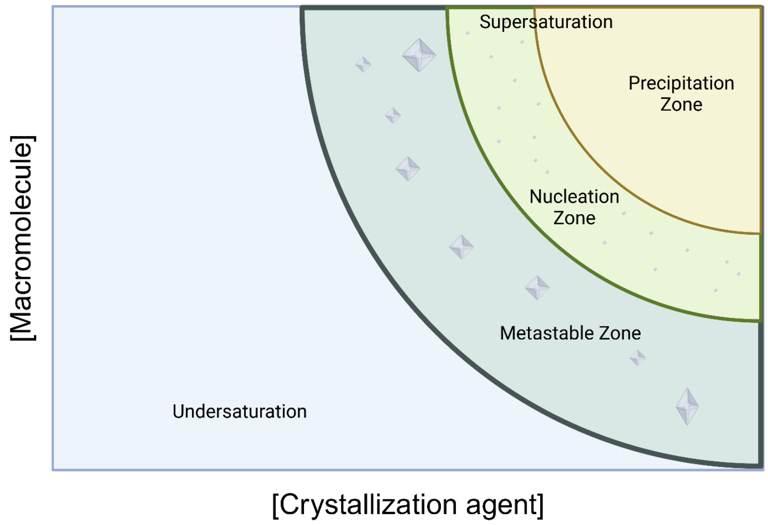


Figure 16. Principle of biomolecule crystallization.
Biomolecule crystallization, and especially proteins, is dictated by chemical, physical and biochemical parameters. The diagram illustrates the dependence of macromolecule and precipitant concentration to induce nucleation points from which a crystal can grow. Increasing too much of those concentrations can lead to precipitation of the sample (supersaturation area).

The most common method for crystallization experiments is vapor diffusion, either by hanging or sitting drop approaches, the latter being the system implemented in the platforms that we accessed for the experimental work of this thesis. We used 96 three-well plates for the screening conditions. These plates are very convenient because we can reduce the amount of protein and ligands used in the screening and also reduce plastic waste. These plates contain three wells for the sample (100 nL) to be tested and a reservoir for the solution containing the precipitant condition, which is different in each position of the 96-well plate. For example, we can screen three protein conditions by varying the protein concentration in each of the three wells, or we can use a single protein concentration and vary the ligand concentration of the ligand itself when analyzing complexes **(Powell, 2021)**.

In both vapor diffusion cases, differences in precipitant concentration cause water molecules to exchange from the drop to the reservoir by evaporation and if the condition is appropriate, the protein or complex can form tiny crystals **(Figure 16)**. Now that we are using synchrotron beam lines to screen the crystals, we harvest the crystals, quickly cryoprotect them, and freeze them in liquid nitrogen. The crystals are stored in a puck and kept under liquid nitrogen until diffraction. For a few years now, we have been able to perform diffraction remotely and collect data while staying in the lab, which is very convenient because it saves time and money on travel and lodging and reduces the impact of the research on the environment.

Before collecting the data, we test the diffraction properties of the crystals by irradiating at several positions in the crystals and identify promising crystals and discard others that diffract badly. The software analyzes the preliminary dataset and suggests the strategy for data collection and experimental setup. The crystal lattice is considered as the symmetrical three-dimensional arrangement of the atoms or molecules inside the crystal. In this context are important concepts such as the lattice points, which are each atom inside the crystal lattice, or the unit cell, which is the minimum repetitive unit in a crystal. The last step is the integration of the intensity of each diffraction spot and determination of the structural factor, a function that explains the amplitude and phase of a wave diffracted from crystal lattice planes. Then, data is scaled and merged. The most used toolbox for X-ray diffraction data processing in most of the synchrotrons is autoPROC, which in addition analyzes the anisotropy of the crystal.

The next step in the process is the calculation of the phase of each diffraction peak, information which is lost during the data acquisition. This can be done mainly by three different methodologies: Multiple isomorphous replacement (MIR), molecular replacement and isomorphous replacement. In this thesis, only molecular replacement was used, which applies a similar biomolecule structure to fit the experimental data. After phase determination, data is further refined in specialized software programs such as CCP4 or Phenix GUI.

In our cases, crystal growth was screened at two temperatures, 4 and 20 °C, at the IBMB-IRB Barcelona Automated Crystallography Platform (PAC).

SMAD4 314-552 constructs R496H was prepared at 4-8 mg/ml in 20 mM Tris pH 7.5, 100 mM NaCl, 2 mM TCEP buffer. Crystals grew at 20 ºC and 4 ºC in different conditions. Best dataset was obtained in 1.4M sodium malonate at pH 7.0. Crystal belongs to the P213 space group and was refined at 2.1 Å resolution.

RREB1 ZF 14-15 was mixed with SerpinE1.12 DNA at 1:1.2 ratio with a final ZF concentration of 5.3 mg/mL in 20 mM Tris pH 7.5, 100 mM NaCl, 2 mM TCEP buffer. Since precipitation was detected upon DNA addition, we increased the NaCl concentration 4-fold. The complex was incubated for at least 30 min at 4 ºC. Crystals were obtained in 25% PEG 3350, 0.2 M sodium chloride, 0.1 M Bis-Tris pH 6.5 after 24 h at 20 ºC. Crystals belong to the C2 space group and the structure was refined at 1.14 Å resolution.

3.11.1 CrystalDirect, Crystallization Information Management System (CRIMS) and PipeDream (EMBL Grenoble)

The High Throughput Facility Lab (HTX), European Molecular Biology Laboratory (EMBL) has developed new technologies and software to handle large numbers of crystals. These include an automated harvester system using a unique crystallization plate format (the CrystalDirect technology) and an interactive platform for easy handling of large data sets, the Crystallographic Information Management System (CRIMS). The additional implementation of automated data processing pipelines is also a key factor **(Cornaciu et al. 2021)**.

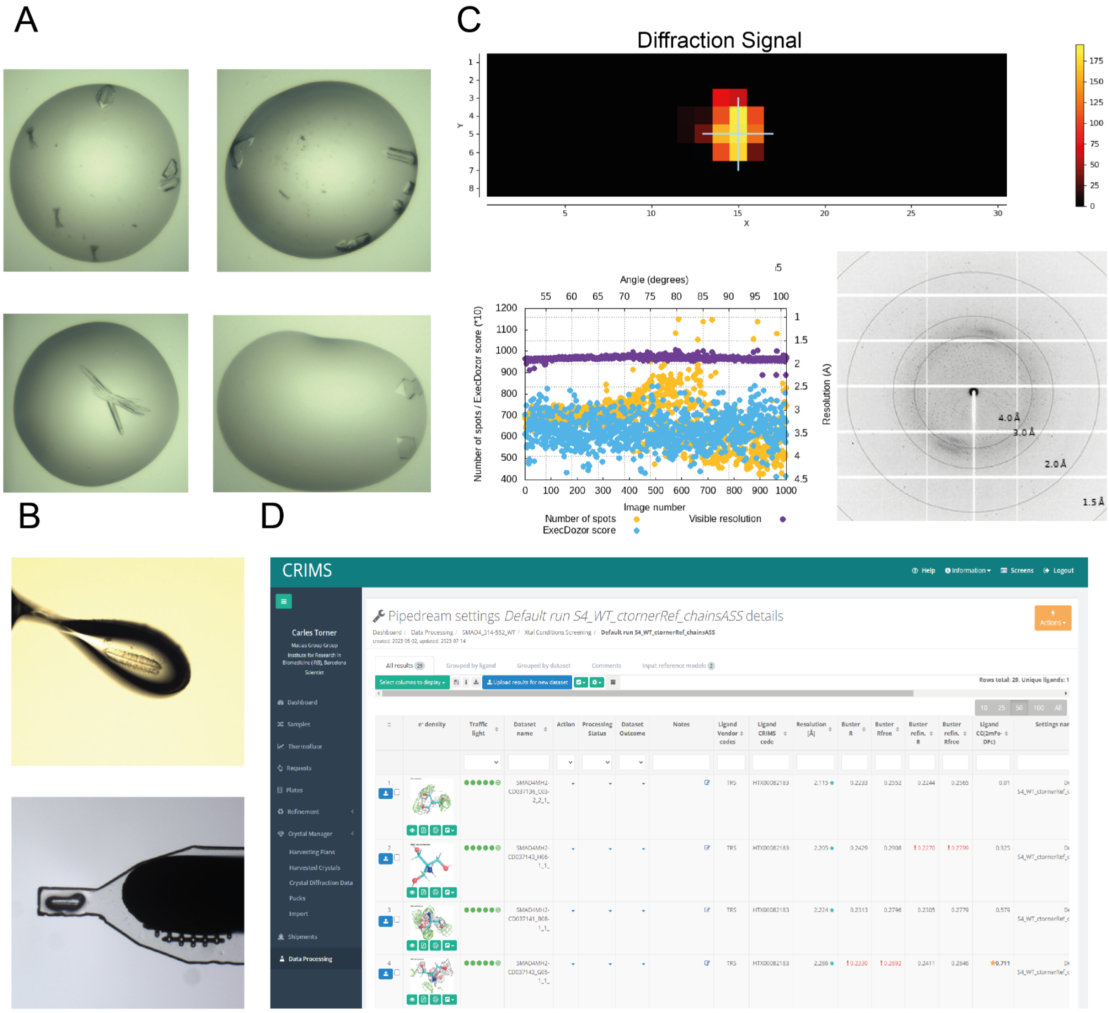
The general approach to fragment or drug screening by X-ray crystallography begins with establishing conditions that yield protein crystals that diffract at 2Å resolution in a highly reproducible manner using a Mosquito robot. Crystallization tests are performed around the selected precipitant solution with the goal of high production of quality crystals in specially designed plates for use with the automated harvester. These plates, called CrystalDirect 2D or CrystalDirect 3D, have a flat and thin plastic layer on the bottom. The crystals are evaluated for diffraction pattern quality and resolution.

The same Mosquito robot is also used to perform a top-drop soak. The goal of this step is to optimize ligand binding without damaging the crystals. One of the conditions optimized is the incubation time of the crystals with the compound dissolved in DMSO. The crystals are harvested using the Crystal Direct Harvester, which uses a laser cutter to cut the plastic of the plate at the region containing the crystal in a desired loop-like shape. This cut region is glued into a plastic holder and automatically stored in a crystallography puck.

Once the diffraction data is obtained, it is transferred to CRIMS where it is automatically processed. The first step is to load a reference model into the platform, including the model and electron density map. Complex data sets are selected and processed through Pipedream, a pipeline developed by Global Phasing Limited (Cambridge, UK), which includes several steps of refinement and docking of the desired compound (those that should be present in the crystallization state or in the soaking experiment).

Initial data processing is performed by autoPROC **(Vonrhein *et al.*, 2011)** and its STARANISO package. Refinement and ligand fitting are then performed by Buster and Rhofit, respectively. Buster is an automated refinement program that uses maximum likelihood and maximum entropy techniques. Rhofit, on the other hand, is a tool for fitting a ligand into an electron difference density map, which can change the length of bonds and angles in the compound of interest. All the steps from the crystallization procedure to data processing are recorded in the CRIMS platform **(Figure 17).**

In our project, we used top-drop soaking, co-crystallization and co-crystallization combined with a back-soaking approach. We performed automatic and manual harvesting and tested conditions with and without cryoprotective agent (CPA). We used 24-26% PEG3350 with 0.2 M $(NH_4)_2SO_{4.}$ Drop volume ranged between 200 and 600 nL, depending on the approach and final ligand concentration was method and DMSO sensitivity dependent. Co-crystalization was performed with a ligand final concentration of 2mM and 2% DMSO. Protein and the ligands were incubated at room temperature for at least half an hour before plate printing. Back soaking was performed to remove the sulfate from the generated crystal. In this approach, crystals were manually harvested and incubated in a drop of a solution containing 20% PEG 3350, 15% glycerol, 10% ethylene glycol, 2 mM ligand and 2% DMSO.



CRIMS

Help · Information · Screens · Logout

Carles Torner
Macias Group Group
Institute for Research in Biomedicine (IRB), Barcelona
Scientist

Dashboard · Samples · Thermofluor · Requests · Plates · Refinement · Crystal Manager · Harvesting Plans · Harvested Crystals · Crystal Diffraction Data · Pucks · Import · Shipments · Data Processing

Pipedream settings *Default run S4_WT_ctornerRef_chainsASS* details

Dashboard / Data Processing / SMAD4_314-552_WT / Xtal Conditions Screening / **Default run S4_WT_ctornerRef_chainsASS**

created: 2023-05-02, updated: 2023-07-14

Actions

All results 29 · Grouped by ligand · Grouped by dataset · Comments · Input reference models 2

Select columns to display · Upload results for new dataset

Rows total: 29. Unique ligands: 1

10 25 50 100 All

::	e⁻ density	Traffic light	Dataset name	Action	Processing Status	Dataset Outcome	Notes	Ligand Vendor codes	Ligand CRIMS code	Resolution [Å]	Buster R	Buster Rfree	Buster refin. R	Buster refin. Rfree	Ligand CC(2mFo-DFc)	Settings nar
1		●●●●●✓	SMAD4MH2-CD037136_C03-2_2_1_					TRS	HTX00082183	2.115 ★	0.2233	0.2552	0.2244	0.2565	0.01	De S4_WT_ctornerRef_c
2		●●●●●✓	SMAD4MH2-CD037143_H06-1_1_					TRS	HTX00082183	2.205 ★	0.2429	0.2908	! 0.2270	! 0.2799	0.325	De S4_WT_ctornerRef_c
3		●●●●●✓	SMAD4MH2-CD037141_B08-1_1_					TRS	HTX00082183	2.224 ★	0.2313	0.2796	0.2305	0.2779	0.579	De S4_WT_ctornerRef_c
4		●●●●●✓	SMAD4MH2-CD037143_G05-1_1_					TRS	HTX00082183	2.286 ★	! 0.2330	! 0.2892	0.2411	0.2846	★0.711	De S4_WT_ctornerRef_c

Figure 17. CRIMS management system.

A. Crystals are observed in several conditions. The information for the condition is available in the system. B. Crystals can be selected for diffraction through the crystal pointing option. C. Crystals can then be harvested manually (top) or with CrystalDirect system (bottom). Harvested crystals are diffracted in the ESRF synchrotron (MASSIF beamline). D. Data can then be processed through Pipedream. Results can be checked in CRIMS, where mtz and pdb files can be downloaded for the density maps and models in each of the refinement steps.

The structures presented in this thesis that were obtained through this pipeline were:

1. CRIMS_Structure 1, SMAD4 314-552 WT-2.115 Å; space group F4132 (cell dimensions: 197.0715, 197.0715, 197.0715, 90.000, 90.000, 90.000); Completeness of 95.7%; R-value of 0.2244, Rfree-value of 0.2565.
2. CRIMS_Structure 2, SMAD4 314-552 WT-2.224 Å; space group F4132 (cell dimensions: 197.1922, 197.1922, 197.1922, 90.000, 90.000, 90.000); Completeness of 93.6%; R-value of 0.2305, Rfree-value of 0.2779.
3. CRIMS_Structure 4, SMAD4 314-552 WT in complex with VP21- 3.039 Å; space group F4132 (cell dimensions: 196.3596, 196.3596, 196.3596, 90.000, 90.000, 90.000); Completeness of 90.8%; R-value of 0.2290, Rfree-value of 0.2581. RSCC: 0.734.

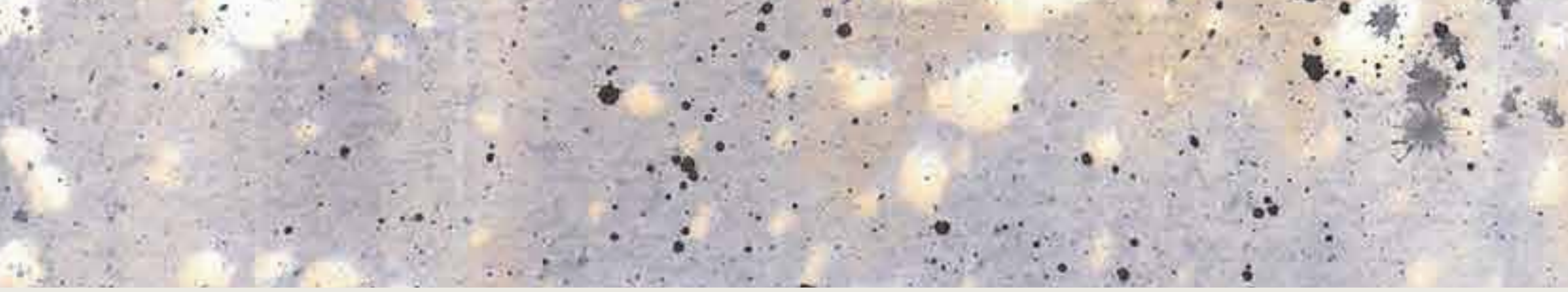
3.12 Small angle X-ray scattering

In a SAXS experiment applied to structural biology, a solution containing biomolecules is irradiated with X-rays produced in a synchrotron to obtain its scattering pattern. Differently as in X-ray crystallography, crystallization of the sample is not needed, allowing the study of the sample in a sort of different conditions and the exploration of changes driven by buffer composition, pH or concentration. The output of the assay is low-resolution data, the product of the scattering of the sample. Scattering detection is produced at small angles relative to the incident beam, and pattern is related to shape and size of the measured particles. The angles of the diffraction then are correlated with the distances between the atoms in the solution and, through solvent subtraction, the analysis can focus on the atomic distances in nanometer scale between the atoms of the biomolecule or complex of interest. The isotropic scattering pattern is recorded by 2D detector and is radially averaged. Results are often illustrated as the scattering intensity (I) dependent on momentum vector (q) **(Da Vela and Svergun, 2020)**.

Different types of analysis can be performed from SAXS data as an assessment of the radius of gyration of the particle (Rg) and the presence of aggregates in the sample (Guinier analysis), the calculation Dmax or maximum distance between two point in the data set (Pair distance distribution function or P(r)) and protein flexibility (Kratky Plots).

SAXS data were acquired on Beamline 29 (BM29) at the European Synchrotron Radiation Facility (ESRF, Grenoble, France). Measurements were performed at 12.5 keV, 100% transmission, low viscosity and 0 s wait time. Data were recorded on a Pilatus 1 M detector, at 10 ˚C. Ten frames per sample were collected for 1 s each. Solvent from each sample elution was collected and their scattering data were acquired to account for buffer contribution. Image conversion to the 1D profile, scaling, buffer subtraction and radiation damage accession was done using the in-house software pipeline available at BM29.The sample buffer optimized for SAXS analysis was 25mM Tris pH 9.0, 100mM NaCl, 2mM TCEP.

Subtracted data was analyzed and compared using ATSAS package in Primus **(Konarev et al., 2003)** or BioXTAS RAW **(Hopkins, Gillilan and Skou, 2017)**.



RESULTS

4.RESULTS

4.1. MyS variants

MyS variants produce an increment of SMAD4 and phospho R-SMADs protein levels in patient cell lines, together with decrease of SMAD4 ubiquitination. These alterations lead to a dysregulation of gene transcription. We have expressed and purified the four described SMAD4 MyS variants (R496C and I500V/M/T). These mutations belong to the gain-of-function class and have been shown to induce a decrease in SMAD4 ubiquitination and an increase in SMADs protein levels in cells **(Le Goff, Michot and Cormier-Daire, 2014)**.

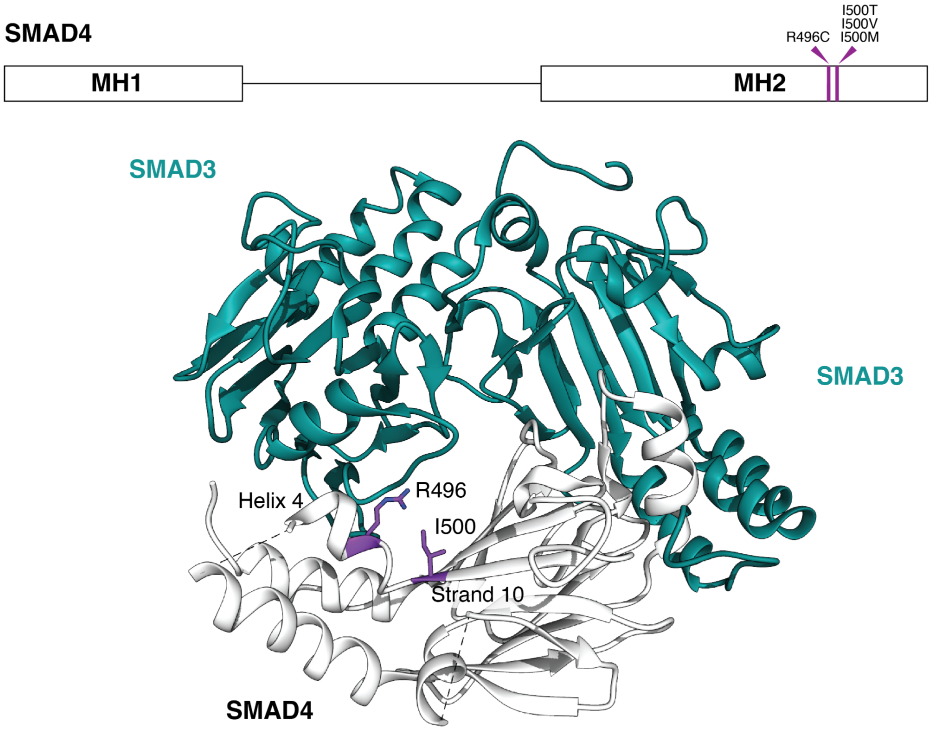


Figure 18. Location of the R496 and I500 residues in SMAD4 MH2 domain.
The positions are indicated in the SMAD4-SMAD3 heterotrimeric complex, PDB code: 1U7V. In the complex, two SMAD3 MH2 domains are shown in green and the single SMAD4 MH2 domain is shown in white. Variants are shown using the ball and stick representation (violet). The dashed lines indicate regions in the structure of SMAD4 that could not be determined from the electron density maps.

Both R496 and I500 residues are localized on the surface of the SMAD4 MH2 domain **(Figure 18)**. In the context of the heterotrimer structure, we noticed that both positions are closed to residues of the R-SMADs, in the binding interface that defines the heterotrimer. We hypothesized that the functional differences described for the variants might correlate with changes in biophysical properties of the variants with respect to the WT protein. To address these questions, we performed a series of *in-vitro* assays using a variety of complementary techniques to measure properties like thermal stability of the mutants and complexes, changes in the affinity for R-SMADs and also changes in the kinetics of the complex formation.

Using several biophysical techniques **(Table 3)** we have observed that the point differences in sequence observed in the variants have little effect on the fold and stability of the SMAD4 MH2 domain. However, the amino acid changes induce an increment on the final amount of the heterotrimeric complexes with R-SMADs, which we could quantify using Mass Photometry. This increment is the consequence of two effects, an increase in affinity for R-SMADs along with an increment of the complex stability, concomitant with a decrease of the dissociation rate of the complexes. Under the conditions studied, heterotrimeric complexes represent 20% of the total number for the WT, and 40 and 50% for the MyS variants, which corresponds to an increase of 2-2.5 times compared to the WT control. MyS variants also have 3-5 times higher affinities for SMAD3 compared to the WT SMAD4 protein, and there is a clear difference in the total number of heterotrimers formed. We also showed that R-SMADs are thermally stabilized upon SMAD4 binding, and that MyS variants produce a significant increase in this stabilization, higher than the WT SMAD4.

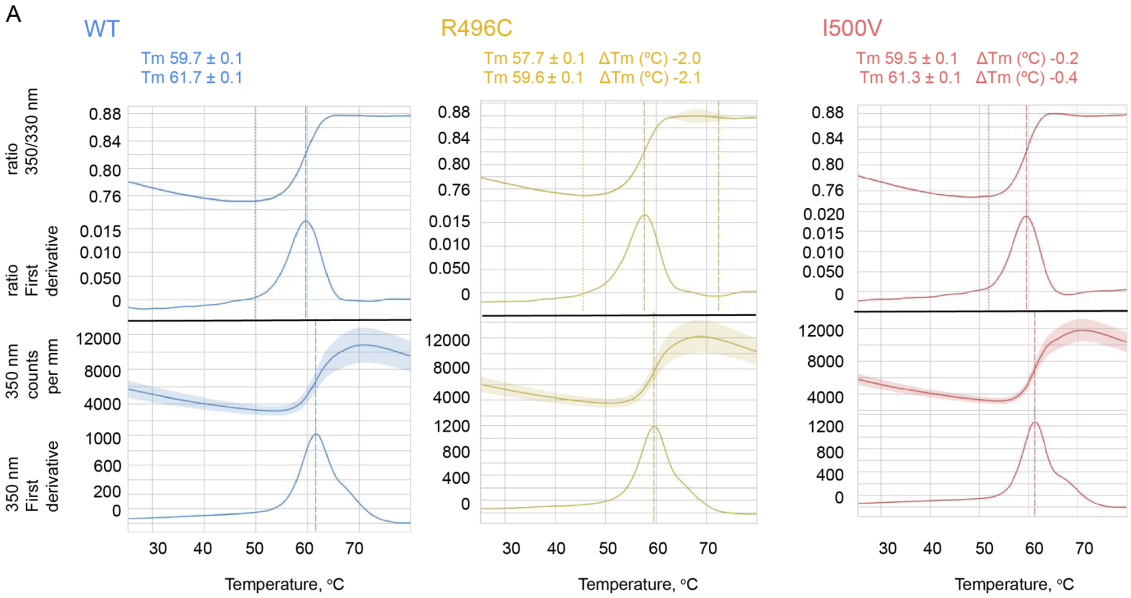
Table 3. Biophysical techniques used in this work.

Technique	Information	Application
DSF and nanoDSF	Thermal stability	Changes in thermal stability in mutants and in complexes
SAXS	Overall fold in solution	Effects of the mutations in the fold and aggregation
SEC-MALS	Particle size	Qualitative effects of the mutations in the association with R-SMADs
Mass photometry	Particle counts	Quantitative effects of the mutations in the association with R-SMADs
SPR	Binding and Kinetics	Effects of the mutations in the folding and dissociation kinetics
ITC	Binding and thermodynamics	Effects of the mutations in the affinity and thermodynamics of the complex
X-Ray crystallography	Structure	Effects of the mutations in the structure

4.1.1. The MyS variants do not affect the stability or the fold of the SMAD4 MH2 domain

4.1.1.1. Biophysical characterization of the MyS variants following changes in the protein stability

Gain-of-function effects are uncommon among SMAD4 variants. Increased SMAD4 protein levels in patient-derived cell lines may be associated with increased thermal stability of SMAD4, resulting in an increased half-life in cells, as previously proposed in **(Le Goff *et al.*, 2011)**. To analyze the effect of the mutations in the protein fold, we first measured the thermal stability of the protein, which is often altered in missense disease associated variants **(Bustad *et al.*, 2013; Fu *et al.*, 2021; Puglisi, 2022)**. Using nanoDSF, we were able to measure the T_m of the SMAD protein variants. NanoDSF has the advantage of not using dyes that may affect protein structure and dynamics, and was used with the two most common variants identified in MyS, R496C and I500V. T_m and ΔT_m values were calculated from unfolding fluorescence signals using the 350/330 nm ratio. The values are almost the same (I500V) or showed a small decrease in T_m (R496C with respect to the WT protein). We also measured the T_m and ΔT_m values for all samples, using conventional DSF **(Figure 19, Supplementary Tables 1-2)**. The results revealed that MyS variants do not have an increased stability compared to the WT SMAD4 domain.



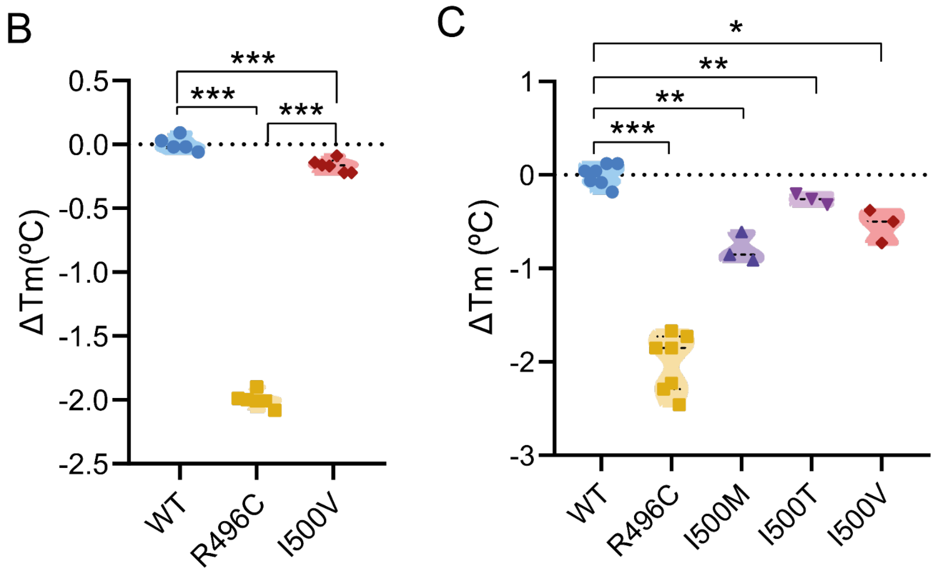


Figure 19. DSF studies of SMAD4 MH2 domains MyS variants

A. Unfolding profiles measured by nanoDSF of SMAD4 272-552 WT (blue), R496C (yellow) and I500V (red). B. Changes in T_m and comparison between values calculated from unfolding fluorescence signals using the 350/330 nm ratio. C. The study was extended to other variants using DSF. The differences observed are measured with respect to the SMAD4 WT and are statistically significant. The differences between the thermal stability of I500V and R496C are also significant.

4.1.1.2. Changes in the overall fold of the domain

We also investigated whether there were structural changes in the MyS variants compared to the WT that could help explain the gain-of-function effects of these amino acid changes, even without an increase in thermal stability. We know that there are regions in the MH2 domain that are not visible in electron density maps due to dynamic properties. Remarkably, these regions are visible in solution using structural biology techniques. We thought that perhaps changes in these or other domain regions could be introduced by the variants.

To test this hypothesis, we chose SAXS as a structural analysis technique because it provides insights into the shape, volume, and flexibility of proteins. This technique does not require highly concentrated samples and, unlike NMR, we do not need to label proteins with specific isotopes. In addition, our lab has recently optimized a pipeline for the analysis of the SMAD4 MH2 WT domain using this technique, meaning that we have a benchmark to compare the effects of the amino acid changes with respect to the WT. Therefore, we used SAXS to compare SMAD4 WT and the I500V MyS variant, which is the most common variant found in MyS patients and is less likely to aggregate than other variants. We measured the SAXS curves at 20, 40 and 80 µM for both proteins, but we had to discard the data sets at 80 µM because we observed protein aggregation in the

Guinier plot and this feature leads to artifacts and misinterpretation in the analysis. In the Kratky plot **(Figure 20)** of the 20 and 40 µM data, we observed that the proteins are composed of rigid and flexible parts and the protein behavior – and the flexibility in particular – depends on the concentration, being this effect more significant at 20 µM. The flexibility is attributed to the presence of the SMAD activator domain (SAD), a flexible region with a long loop that precedes the compact core of the MH2 domain. We believe that the dynamic properties of this region are affected by a compaction effect associated with increasing concentration.

However, we noticed that the comparison of the Kratky plots of the WT and the variant did not show any major changes in domain flexibility, suggesting that the variant has a predominant conformation for both the SAD and the core structure that is very similar to the WT domain **(Figure 21)**. The similarity was confirmed by the analysis of the atomic distance distribution – P(r) function – and the average maximum distance between the atoms of the particles in solution –Dmax–, which gave almost identical values (WT :10.5 nm at 20 µM and 10.0 nm at 40 µM, I500V, 10.5 and 9.8 nm) for both 20 µM and 40 µM samples (**Figure 21)**. The slow decay of the P(r) function in the 20 and 40 µM samples for both constructs can indicate either aggregation or non-globularity. However, in these cases, the analysis of the Guinier plots allowed us to discard the presence of aggregation.

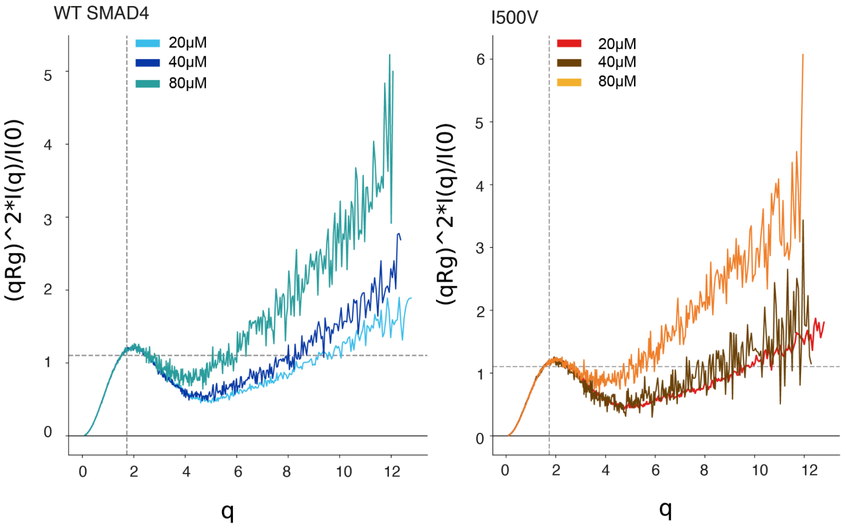


Figure 20. SAXS data analysis of SMAD4 MH2 domain WT and variants.
Kratky plots for the WT (left) and I500V variant (right) at 20, 40 and 80 µM.

Overall, from the thermal stability and the SAXS analyses, we conclude that there are no observable differences in the fold of WT and the I500V variant. We then hypothesized that the effect of the amino acid change and the increased stability of the MyS variants reported in the literature might be related to changes in the oligomerization equilibrium of SMAD4 and R-SMADS rather than intrinsic changes in the isolated SMAD4 MH2 domain.

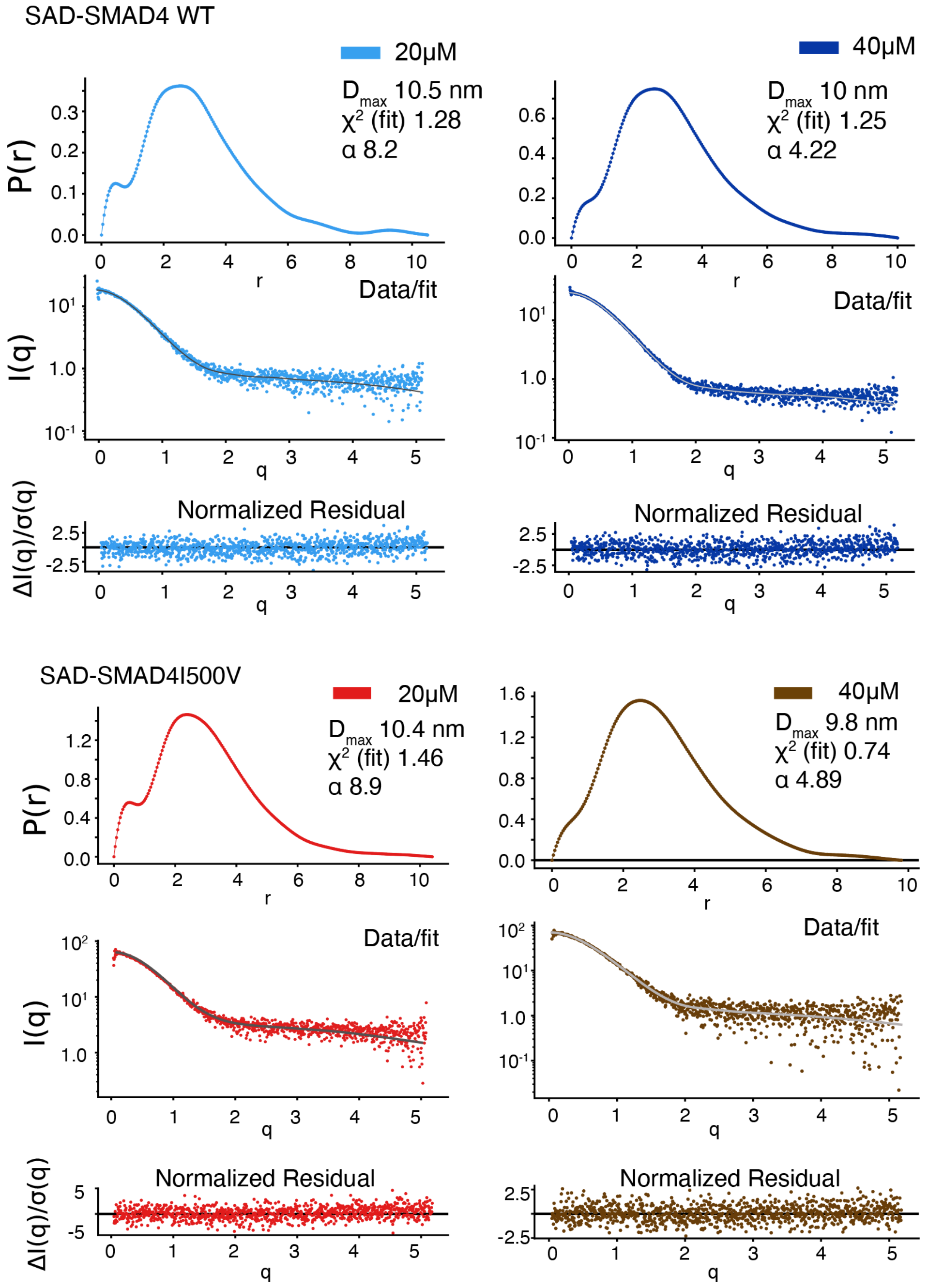


Figure 21. SAXS distance distribution analysis of SMAD4.

Pair distance distribution (P(r)) and scatter intensity fit for SMAD4 WT and I500V variant. The curves are very similar for the WT and I500V variants. The left shoulder visible at the lowest concentration is characteristic of the compact part of the domain, while the second maxima correspond to the extended and flexible regions of the protein construct.

4.1.2. Myhre syndrome variants form more stable complexes with R-SMADs than the WT SMAD4 protein

The results of the basic characterization of the domain stability and conformation in the MyS context inspired us to focus our study on PPIs variation **(Figure 22)**.

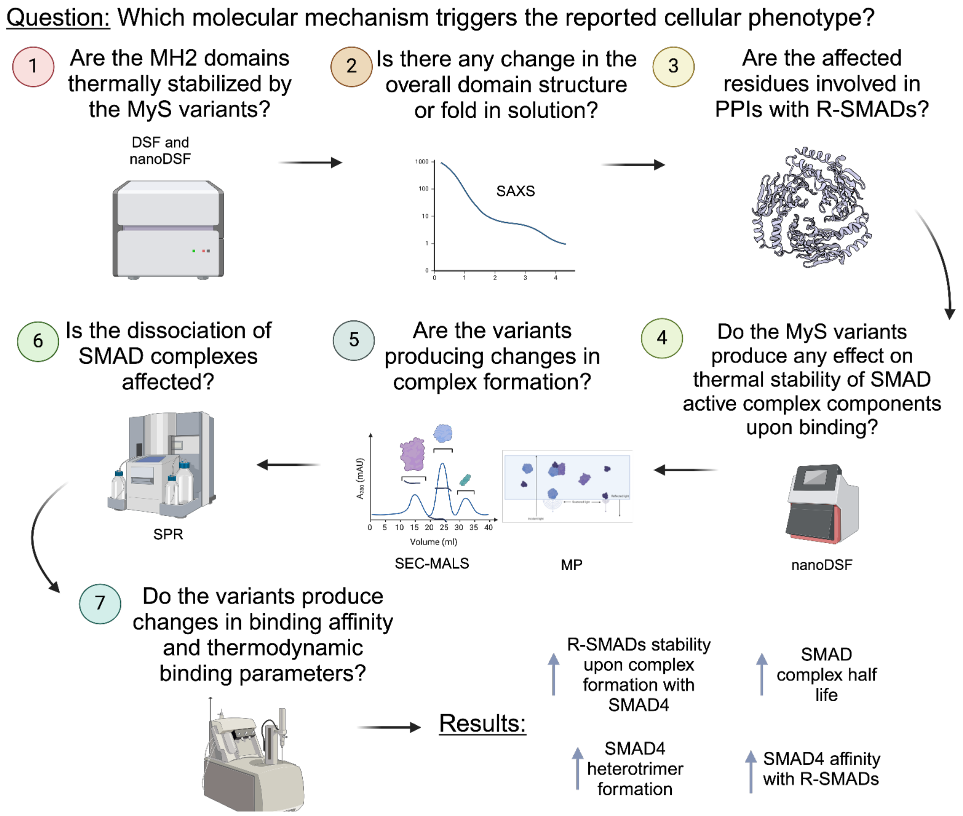

Question: Which molecular mechanism triggers the reported cellular phenotype?
1 Are the MH2 domains thermally stabilized by the MyS variants?
DSF and nanoDSF
2 Is there any change in the overall domain structure or fold in solution?
SAXS
3 Are the affected residues involved in PPIs with R-SMADs?
4 Do the MyS variants produce any effect on thermal stability of SMAD active complex components upon binding?
nanoDSF
5 Are the variants producing changes in complex formation?
SEC-MALS
MP
Volume (ml)
6 Is the dissociation of SMAD complexes affected?
SPR
7 Do the variants produce changes in binding affinity and thermodynamic binding parameters?
Results:
R-SMADs stability upon complex formation with SMAD4
SMAD complex half life
SMAD4 heterotrimer formation
SMAD4 affinity with R-SMADs

Figure 22. Schematic representation of the hypotheses and conclusions.

We measured the effects in the T_m of the R-SMADS MH2 domain (SMAD1 and SMAD3, examples of BMP and TGF-β receptor activated SMADs) in the presence of increasing amounts of SMAD4 MH2 domains using nanoDSF, either WT or MyS variants. We also used the R361G mutant as a negative control in the experimental setting, given that this point mutation has been described in cancer and is known to prevent heterotrimer formation. As for the R-SMADs, we have used two mimics of the phosphorylation state, SMAD3-DVD and SMAD1-EEE respectively. SMAD3 forms homo-dimers and homotrimers in the absence of SMAD4 **(Gomes *et al.*, 2021)** in solution, whereas SMAD4 is mostly monomeric. The melting temperature of isolated SMAD4 and R-SMADS is quite different (almost 12 degrees higher in SMAD4 than in SMAD3) and their unfolding appears as separate events in 350/330 nm fluorescence ratio representation **(Supplementary Figure 1-3)**. We believe that this large difference in T_m has biological significance and may promote the interaction of R-SMADs with SMAD4 in native contexts, favoring the formation of heterotrimeric rather than homotrimeric forms.

We observed that the addition of either WT or MyS SMAD4 variants produced an increase in T_m of both SMAD3 and SMAD1, whereas the addition of the R361G SMAD4 variant did not have an effect, as expected for a cancer variant. **(Chacko *et al.*, 2004) (Figure 23, Supplementary Table 3)**. The effect of adding SMAD4 to SMAD3 as a change in its stabilization starts at 29 µM SMAD4 concentration. This effect is more pronounced for the MyS variants than for the WT SMAD4. Moreover, while the effect can be observed for both SMAD3 and SMAD1, it is more pronounced in the case of SMAD1, and at lower SMAD4 concentration.

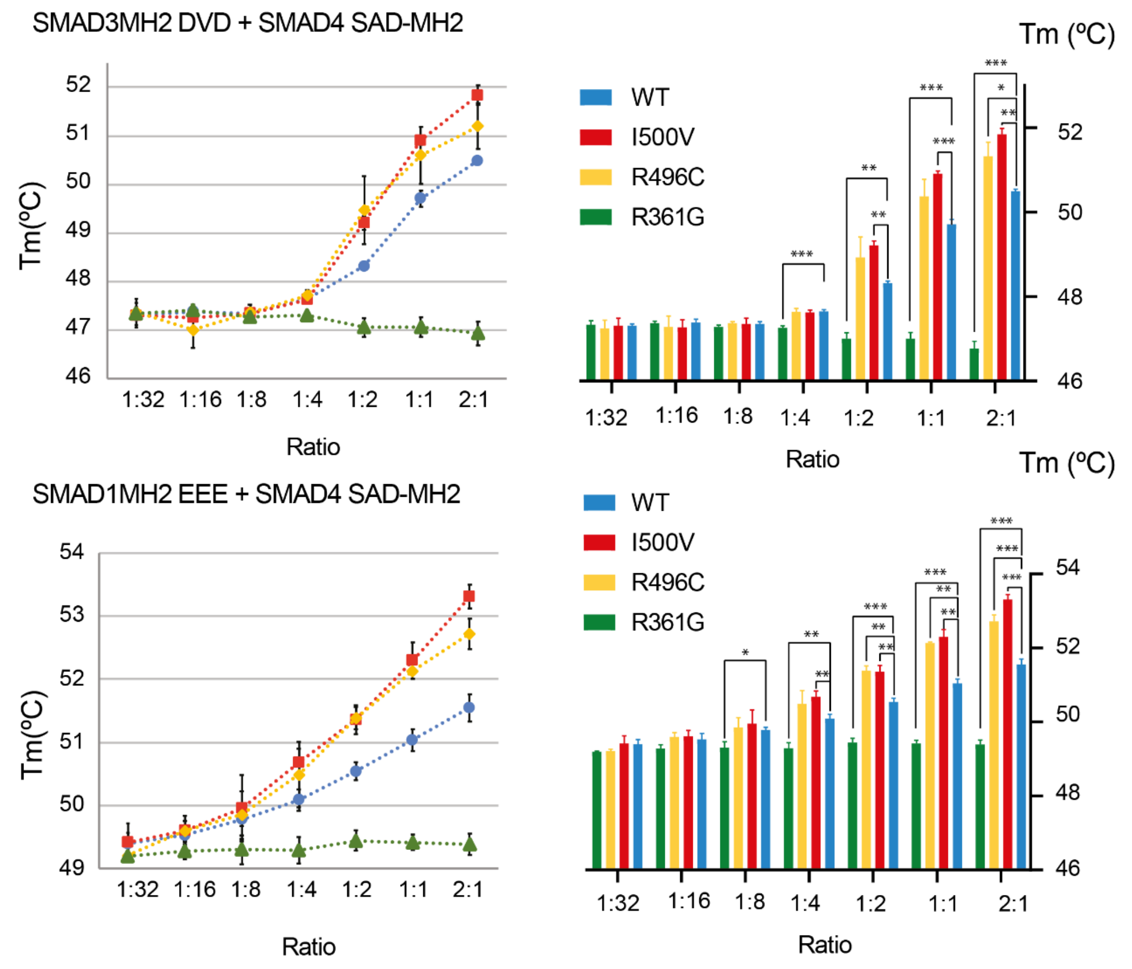


Figure 23. R-SMADs thermal stabilization in the presence of SMAD4 WT and variants measured by nanoDSF. The differences observed are measured with respect to the SMAD4 WT and are statistically significant.

We have applied two different biophysical strategies to visualize the interactions between the MH2 domains, for WT and mutant proteins: Size Exclusion Chromatography with Multi-Angle Light Scattering (SEC-MALS) and Mass Photometry (MP). SEC-MALS allows us to determine the average mass of the particles for each peak of a size-exclusion chromatography separation. From this value, we can calculate the composition of the complexes. As SMAD4 and SMAD3 MH2 domains have a very similar mass, we used a SUMO-SMAD4 ($MW_{theorical}$=38.25 kDa) construct to increase its mass, whereas the SMAD3 is native ($MW_{theorical}$=26.88 kDa).

To test the behavior of the samples in the experimental conditions, SMAD4 and SMAD3 controls were injected independently in a SEC-MALS system **(Figure 24A,B)**. The results show negligible levels of aggregation and high purity of the protein species. We also observed that SMAD4 MyS variants and WT protein behave as monomers, whereas SMAD3 (189-425 DVD construct) elutes as a single peak containing a mixture of oligomeric states **(Figure 24A,B)**.

When we injected a mixture of SMAD4 and SMAD3, with SMAD4 being in excess, we observed the formation of a complex, whose apparent mass is larger than that of the oligomers of SMAD3 alone, which we interpret as a hetero-oligomer complex. This observation happened for both the WT and the MyS variants, but it is especially noticeable for the latter **(Figure 24C,D)**. As SMAD4 is in excess, the peak for the free monomeric SMAD4 can also be observed.

The results observed by SEC-MALS were corroborated, and the resolution was increased by MP experiments that were performed in a One MP system (Refeyn) at the SPC facility. EMBL Hamburg. MP is a light scattering-based technique that detects single, unlabeled molecules in dilute solutions.

This technique can accurately measure molecular masses in the range of 40 kDa to 5 MDa. The most notable feature of the technique is that it can provide information on the relative abundance of species by molecular counting (Cole et al. 2017). The most important aspect in this project is that the technique can reveal the relative abundances of different biomolecules and their complexes in mixtures at the single molecule level, as well as the complex stoichiometries. The molecular counting is achieved through the quantification of the light scattering of single proteins upon binding to an illuminated glass-water interface **(Soltermann *et al.*, 2020)**. Given this level of accuracy, in MP experiments, buffer conditions need to be optimized to minimize the background counts. This step is particularly challenging when working with "low molecular weight" biomolecules, as buffer noise can make it difficult to properly analyze the sample. After testing several buffers, we found a combination that gave no contamination in the mass range of interest for the SMAD protein system.

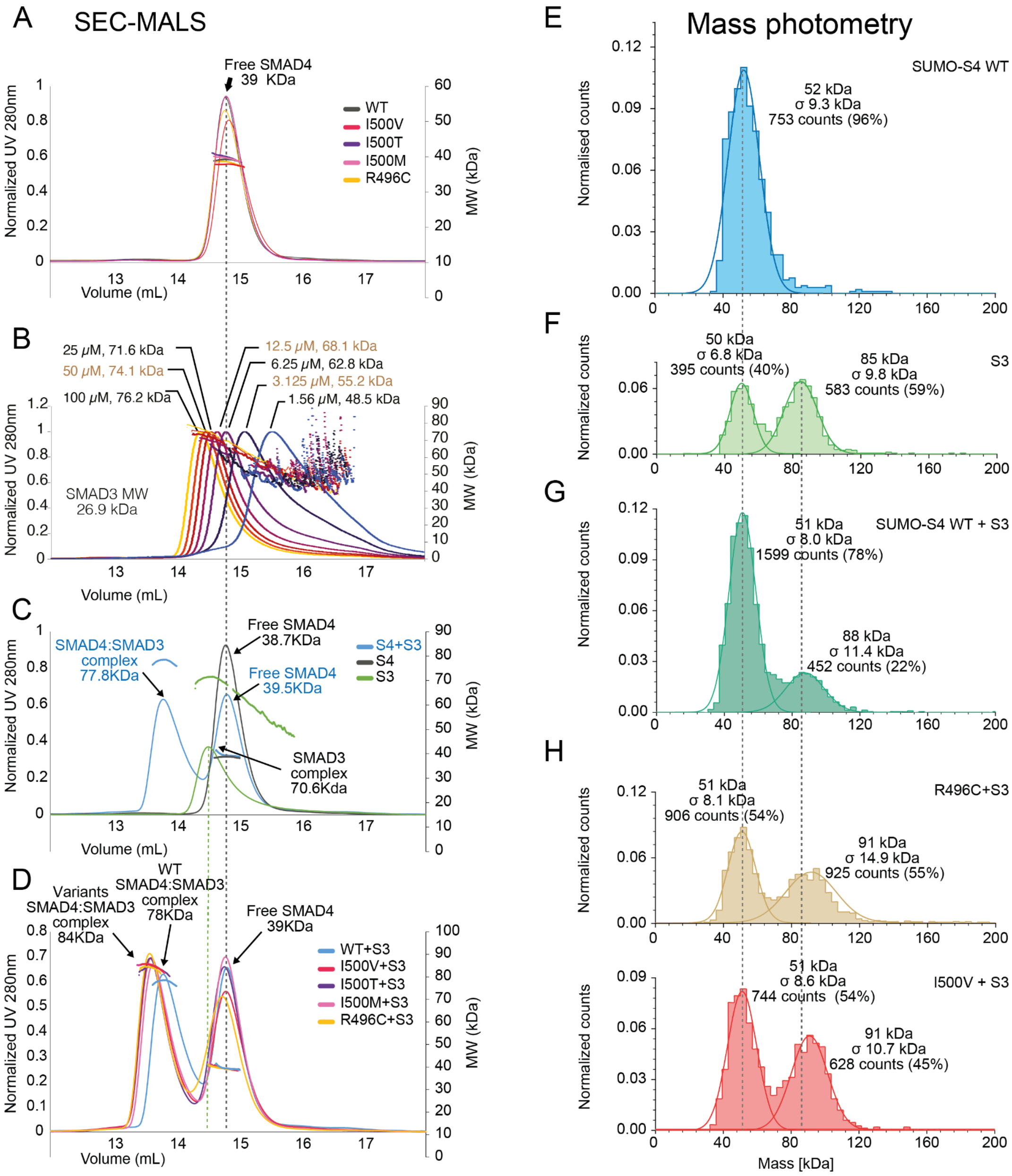


Figure 24. Complexes of WT and MyS SMAD4 variants with SMAD3.

A. Size-exclusion chromatography coupled to multi-angle light scattering detector (SEC-MALS) data for SMAD4 WT and MyS variants show a single peak of a MW that corresponds to a monomer. B. Data corresponding to SMAD3 indicates the presence of dimers and trimers and the absence of monomers even at the lowest concentration. C. The complex between SMAD4 and SMAD3 shows the heterotrimer complex as well as a peak corresponding to the excess of unbound SMAD4. D. Compared to the WT protein, the MyS variants elute as a peak with a higher average molecular weight, indicating that the complex equilibrium is shifted toward the hetero-trimer formation. Mass photometry (MP) was used to quantify the number of particles corresponding to heterotrimeric complexes (left). E-F. Controls of the SMAD4 and SMAD3 proteins in the free state. G. A mixture of SMAD4 and SMAD3 reveals the presence of homo and heterotrimers. H. Complexes with MyS variants show two times more heterotrimer particles compared to the WT scenario (G). Final protein concentrations in MP assays were 5.26 nM (SMAD3) and 2.63 nM (SMAD4). The monomeric constructs used in SEC-MALS and MP experiments had a molecular weight of 38.25 kDa for SUMO-SMAD4 314-552 and 26.88 kDa for SMAD3 189-425 DVD. The SMAD4:SMAD3:SMAD3 heterotrimer weighs 92 kDa, while the SMAD3 homotrimer weighs 80.6 kDa.

Again, in these experimental conditions, SUMO-SMAD4 314-552 constructs (WT and MyS variants) behave as a single peak with negligible aggregation signal **(Figure 24E)**.

Since we were working close to the resolution level of the instrument at low molecular weights, the MW determined by this system was ~50 kDa instead of the expected 38 kDa, but still good enough to prove its monomeric state. With respect to SMAD3, we detected trimeric species, with accurate masses, since this size folds into the resolution range of the system **(Figure 24F)**. For the complexes, we worked at saturating conditions with a 1:2 ratio (SMAD4:SMAD3).

The most relevant aspect of this technique is that we could accurately detect how abundant a given complex is in the measured solution. This complex counting allowed us to detect an increment of heterotrimeric complexes ($MW_{theorical}$=92 kDa) for the MyS variants, as the shape and size of SMAD3 homotrimers and SMAD4-SMAD3 heterotrimers are easily distinguishable. To illustrate these observations, examples of both histograms and Gaussian distributions are superimposed and compared in **Figure 24E**. As depicted, the MW for the heterotrimers were accurately measured, obtaining similar values to the theoretically calculated.

Remarkably, when comparing WT and MyS proteins, there is a clear difference in the total number of heterotrimers formed. Under the conditions studied, heterotrimeric complexes represent 20% of the total number in the WT scenario, as opposed to 40 and 50% for the MyS variants **(Figure 24G,H)**. This implies an increase of 2-2.5 times more heterotrimeric complexes compared to the WT control in the conditions studied. This observation is consistent with the SEC-MALS results previously obtained, and suggests that MyS variants form more stable complexes with different dissociation rates (k_{off}) than the WT.

To complement these observations, we measured the differences in complex dissociation rate through Surface Plasmon Resonance (SPR) and SMAD2/SMAD3 as analytes. The experiments were performed by immobilizing SMAD4 WT, I500V and R496C variants in different channels of the same chip in a covalent manner through amine coupling. Other methods were tested, including non-covalent HisTag or StrepTag capture but, they were discarded due to either non-specific binding of SMAD proteins to the nickel-loaded chip surface (His-tag) or non-specific reactions on the streptavidin-loaded CM5 Cytiva chips. Using the amine coupling immobilization approach, the SMAD4 variants and WT domain were titrated with increasing concentrations of SMAD3 and SMAD2. In these experiments, we observed that SMAD2 binds more tightly to the R496C variant compared to the I500V one, while for SMAD3, the complexes seem to behave similarly. We observed a 5 to 6-fold increase in the half-life time of SMAD4 mutated complexes compared to the WT counterpart **(Figure 25).** In all cases, low χ^2 values indicate a good fit of the data into the 1:1 dissociation model.

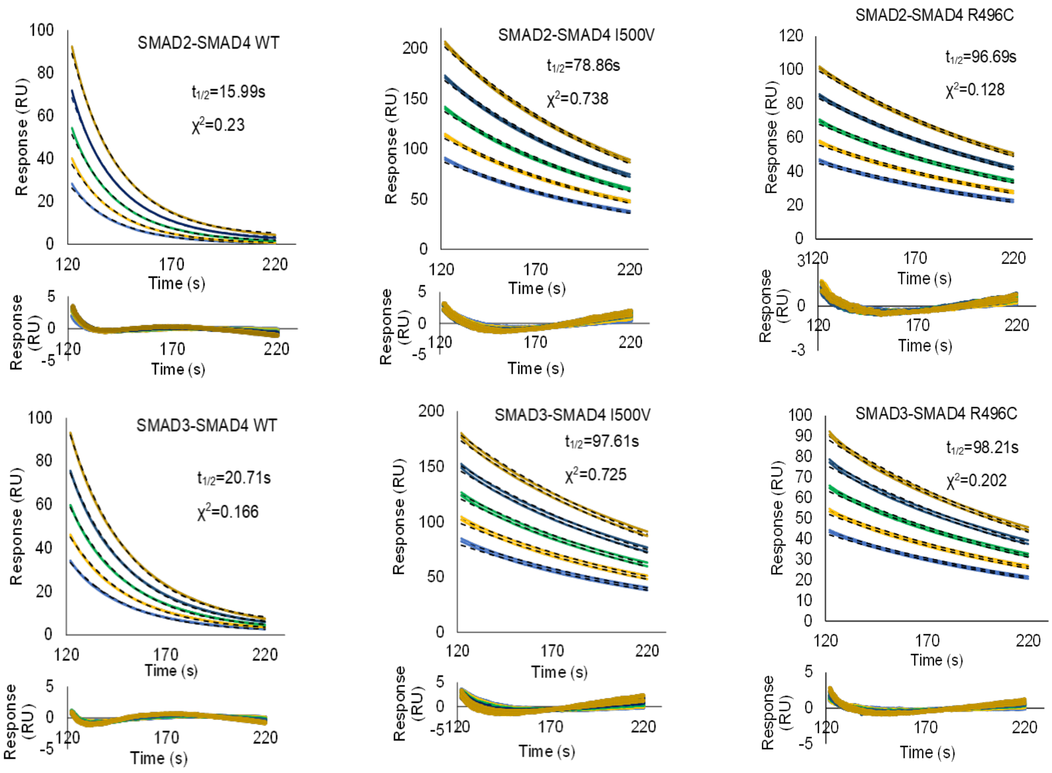


Figure 25. Kinetic and thermal stabilization in the SMAD4 complex with R-SMADs in WT and MyS variants. Analysis of the dissociation phase of SMAD4-R-SMADs MH2 domains' interactions on a Biacore T200. SMAD4 272-552 variants were immobilized in the chip surface while titrating increasing concentrations of activated SMAD2/3 MH2 domains. Analyte concentrations are 1.56, 3.13, 6.25, 12.5 and 25 μM from bottom to top.

4.1.2.1. Affinity of the complexes

We used ITC to quantify the affinity of SMAD3 for SMAD4 WT and variants, using a similar approach to that described in the literature for the WT protein **(Chacko *et al.*, 2004)**. In this work, the authors obtained values of K_D of 58 nM for SMAD4/SMAD3 and 296 nM for SMAD4/SMAD2. We obtained values with about 10-fold change for the SMAD4/SMAD3 WT, 528 nM. We associate this phenomenon with our differences in the proteins constructs and the usage of phosphomimics. Our experiments with the MyS variants showed 3-5 times higher affinities for SMAD3 compared to the WT SMAD4 protein **(Figure 26A)**. The measured stoichiometry (n) is approximately 0.5 which is consistent with the formation of a SMAD3:SMAD3:SMAD4 trimer.

The ITC derived thermodynamic binding parameters **(Figure 26B, Supplementary Table 4)**, revealed similar overall ΔG in all cases, but in the case of the MyS variants and with respect to the WT values, the entropic contribution (-TΔS) is higher, whereas the enthalpic one (ΔH) is lower, suggesting differences in binding modes between SMAD variants. In ligand binding, entropy-enthalpy compensation generally means that a ligand or receptor modification results in a change in the enthalpic contribution to binding that is compensated (off-set) by a similar change in the entropic component of binding. In this case, the increase in the enthalpic contribution of the MyS variants could be due to an increase in hydrogen bond formation, van der Waals interactions, or the strengthening of pre-existing contacts in the complex, supporting the hypothesis that the MyS variants form more stable and long-lasting complexes with R-SMAD proteins than the WT counterpart protein.

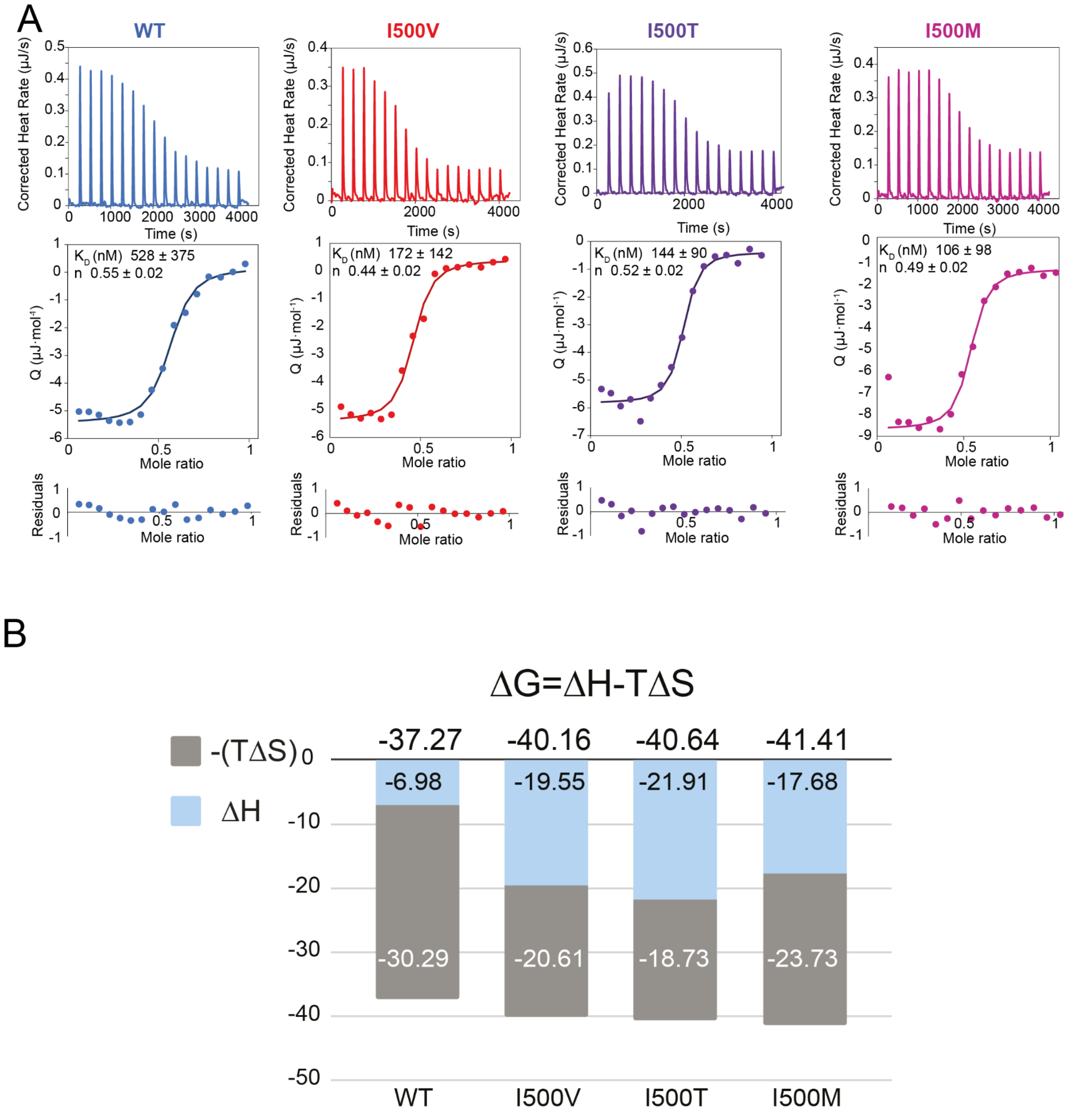


Figure 26. ITC measurement for WT and I500 variants reported in MyS.

A. Thermograms (upper panels) and binding isotherms (lower panels). SMAD3 189-425 DVD was applied to the cell and SMAD4 314-552 was injected. Measurements were performed at 37ºC and at 260 rpm. Fittings were performed using the independent binding site model. B. Thermodynamic ΔG, ΔH and -TΔS parameters derived from ITC, in kJ/mol.

4.2 SMAD4 variants identified in cancer patients and other rare diseases

Missense mutations are the most common alterations of *SMAD4* gene. These mutations, which are often found in the MH2 domain, lead to loss-of-function effects, like the loss of R-SMAD binding capability. In the *SMAD4* gene, there are numerous missense and nonsense mutations. The first case corresponds to a change in the amino acid, and the second type corresponds to truncations that produce a shorter protein.

Although the mutations are distributed along the SMAD4 gene sequence, most of the missense mutations concentrated in the MH2 domain, and in the binding interface with R-SMADs **(Figure 27)**. Some residues present several mutations, leading to different variants, with position R361 being the most commonly mutated site in SMAD4 protein (23% of SMAD4 cancer-related mutations). We also noticed that several oncogenic mutations are localized in the same region as those reported in variants associated with MyS or with other rare diseases as JPS and HHT.

This chapter will describe how different mutations detected in cancer can affect interactions of SMAD4 with R-SMADs. For the analysis we selected 8 different variants: D351G, P356L, R361G, G386D, A406T, K428T, R496H and R515T. Some of the selected cancer-related variants also appear in JPS (G386D and R361G) and R361 in HHT, although in this case, the reported variant is R361C **(Cao, Plazzer and Macrae, 2023)**.

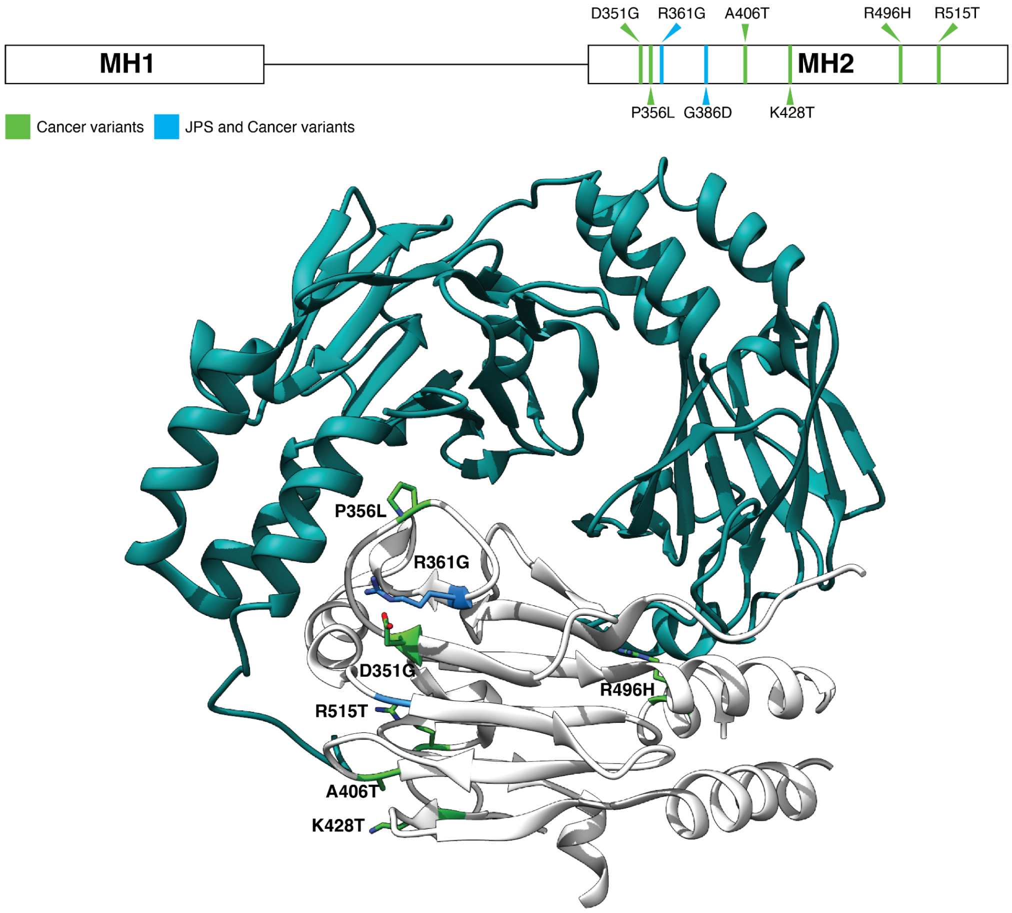


Figure 27. Selected cancer and JPS variants in SMAD4 MH2 domain.
Specific residues and changes are indicated.

We investigated the effects of the residue change on the protein stability and observed that, in contrast with the negligible effect of MyS variants, these changes do not share a common effect on the thermal stability of the MH2 domain of SMAD4. We also observed differences in the effect of the mutations on the complex affinities. Using these affinities, the mutations can be classified into four groups **(Table 4)**.

Table 4. Effect of studied SMAD4 cancer variants in complex formation with SMAD3

Group	Effect of variant in complex formation	Variants
I	No change	A406T, K428T, R515T
II	Decreased affinity	G386D, R496H
III	No complex formation	R361G, P356L, D351G
IV	Increased affinity	MyS mutations

Variants belonging to group IV enhance SMAD complex formation, and we selected the variants at position R496, which are reported in cancer and MyS, for structural analysis.

4.2.1 Effect of amino acid changes in SMAD4 MH2 domain associated with cancer/JPS

As we have done before with MyS variants, we have used DSF analysis to determine whether the variants affect the thermal stability of the protein **(Figure 28, Supplementary Table 5)**. The experiments revealed that some mutations produce changes in T_m compared to the WT protein, while others do not.

P356L, R361G, G386D, A406T and R496H variants displayed a decrease in T_m, and this effect was especially noticeable for G386D, with a ΔT_m ≈-14ºC. Only the R515T variant produced a stabilization of the MH2 domain of SMAD4 (ΔT_m ≈+2ºC) and the rest of the variants were neutral with respect to the T_m. From this data, we conclude that cancer-related mutations do not share a common effect on the thermal stability of the MH2 domain of SMAD4.

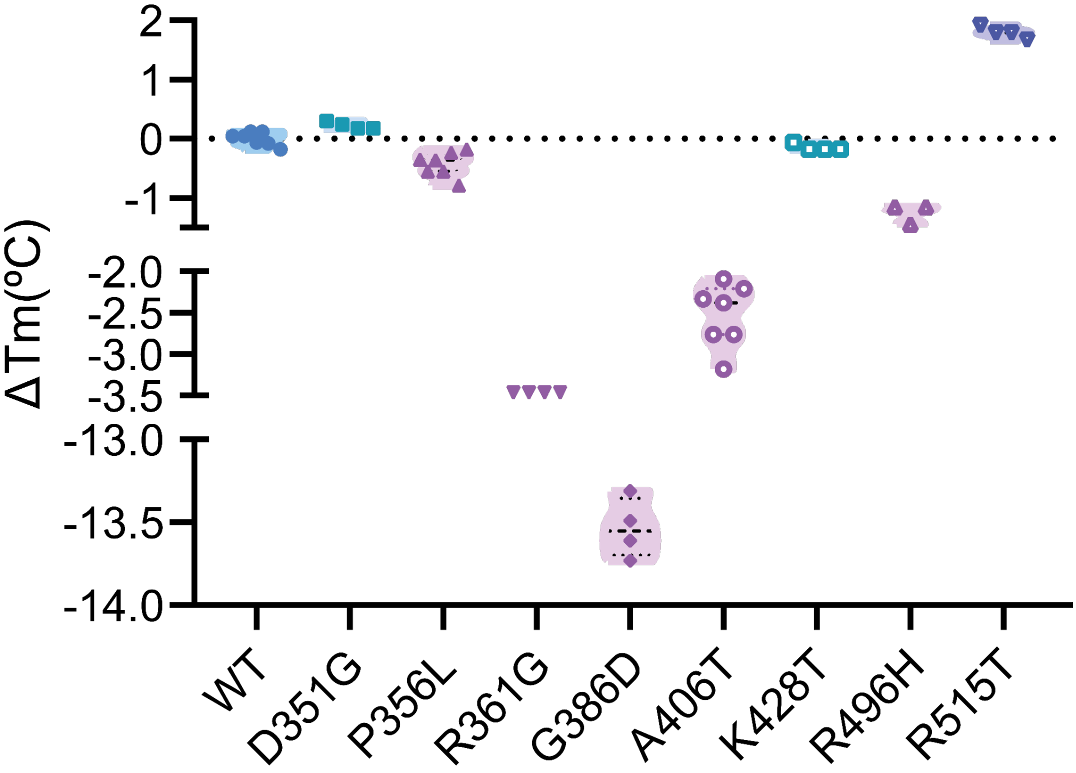


Figure 28. Determination of ΔT_m using DSF.

WT is shown in blue, and the 8 variants in orange. The differences observed are measured with respect to the SMAD4 WT were studied using the Welch's t-test and shown to be statistically significant in all the cases.

4.2.2 Oligomerization properties of the variants with cancer-associated mutations

We hypothesize that, as in MyS variants, the selected cancer-related mutations could affect the dimerization and trimerization of SMAD4 with the R-SMADS. SEC-MALS profiles indicate that all the studied SMAD4 MH2 domains behave as monomers in solution, except A406T, which showed an MW higher than that of a monomer and might indicate a different behavior **(Figure 29)**. In the presence of SMAD3, the profiles indicate that not all cancer variants form the same complexes. In these experiments, the SMAD4:SMAD3 ratio was kept at 2:1 (excess of SMAD4 protein).

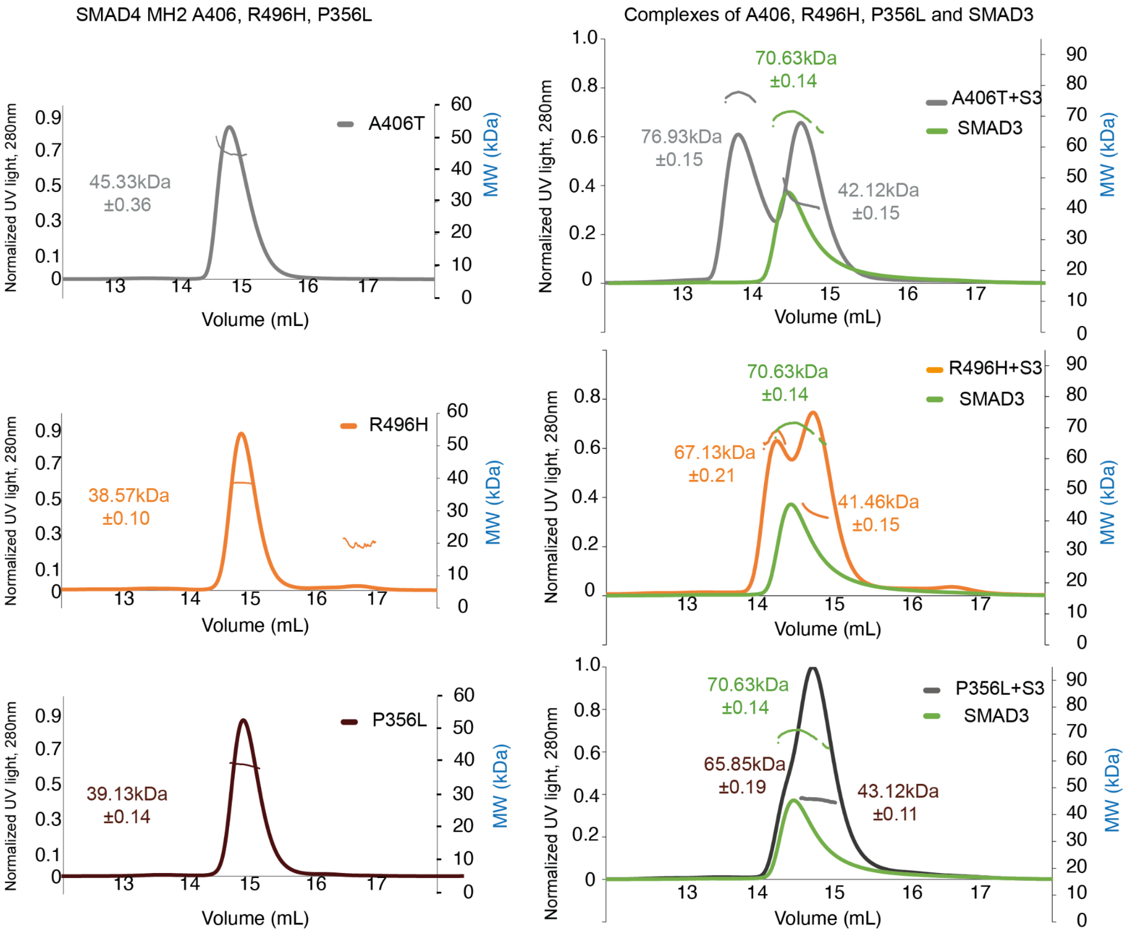


Figure 29. SMAD4 complex formation analyzed using SEC-MALS.

(left) Profiles corresponding to three variants, A406T, R496H and P356L. (Right) Complexes of these variants and SMAD3. SMAD4 samples were measured at 50 µM and SMAD3 at 25 µM.

The interaction of SMAD4 A406T with SMAD3 was further analyzed by ITC to detect if there was a change in the binding affinity **(Figure 30A)**. We observed a slight decrease in the affinity with respect to the WT SMAD4 MH2 domain (K_D=2.61 µM vs. 0.55 µM). While the change is small, we also noticed a shift in the entropy and enthalpy contributions to the binding as well as the entropy-enthalpy compensation effect as in the MyS variants previously analyzed **(Figure 30B)**. This effect could again indicate a change in the mechanism of complex formation, which we plan to analyze in more detail in the future.

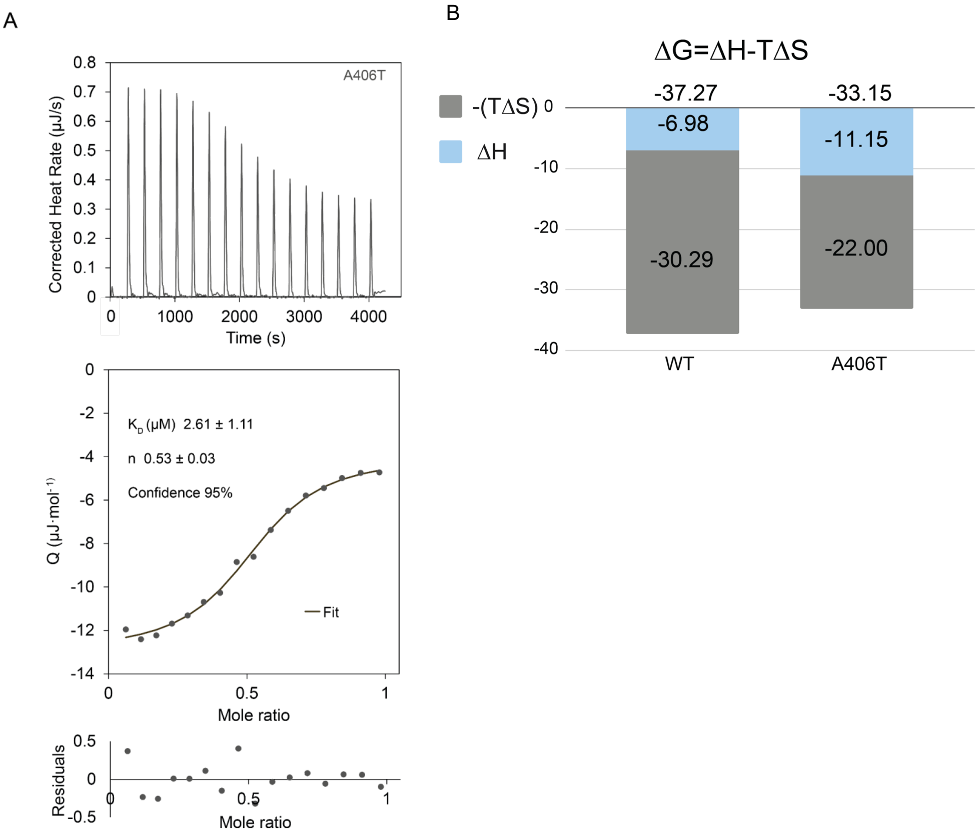


Figure 30. ITC analysis of the interaction of the SMAD4 A406T variant with SMAD3.

A. ITC curves. SMAD3 189-425 DVD was in the cell and SMAD4 314-552 A406T was injected. Measurements were performed at 37 ºC and 260 rpm. Fitting was performed using the independent binding site model. B. Thermodynamic ΔG, ΔH and -TΔS parameters derived from ITC, in kJ/mol.

Using the affinity of the complex as a classifier, the mutations can be separated into four groups. At a given concentration, variants that belong to Group I are those that form complexes with SMAD3 as the WT protein. Group II variants form complexes that elute in SEC-MALS with a lower molecular weight compared to WT complexes, indicating the presence of lower MW associations, for instance as dimers. Group III comprises mutations unable to form SMAD4-SMAD3 complexes. This group includes highly frequent cancer-related mutations. Group IV includes mutations that enhance SMAD4 complex formation with R-SMADs. This includes MyS variants, one of which is also reported in cancer (R496C), although with low frequency.

In summary, as with the thermal stability assays, we have observed that cancer-related mutations exhibit a wide range of responses in terms of complex formation. The majority of the mutations studied show a reduction or direct inhibition of complex formation.

4.2.3 SMAD4 R496H structure, an example of Group II mutation

The residue R496 is mutated in both MyS (R496C) and cancer (R496H). We selected this position for structural studies using X-ray crystallography to explore the effects of the changes in the MH2 domain fold. We got quality diffracting crystals of the R496H variant. In the structures, we observed that the main fold of the MH2 domain is conserved. We also observed that long loops are not visible in the electron density map, as it happens very often in structures of the SMAD4 MH2 domain **(Figure 31A)**. The most notable difference with respect to the WT protein is in the H4 helix, which is shorter and less visible in crystals **(Figure 31A,B)**. The reduced length of this helix is likely to induce changes in the packing of the other two helices that form the three-helix bundle, thus affecting the length of the other helices as well. Moreover, when the mutant structure is compared with the WT structure, we observed that whereas in the WT protein, the Arg residue can form a hydrogen bond with D493, this possibility is absent in this mutant and probably also in the R496C one **(Figure 31C,D)**. The interaction between the R496 and the D493 plays several roles. One is stabilizing the secondary structure of the H4 helix. The second is that the Arg-Asp interaction orients the Asp side chain for binding to SMAD3 when the heteromeric complexes are formed. The combination of these two effects might explain the decrease in thermal stability and complex formation we observed for the R496H mutant.

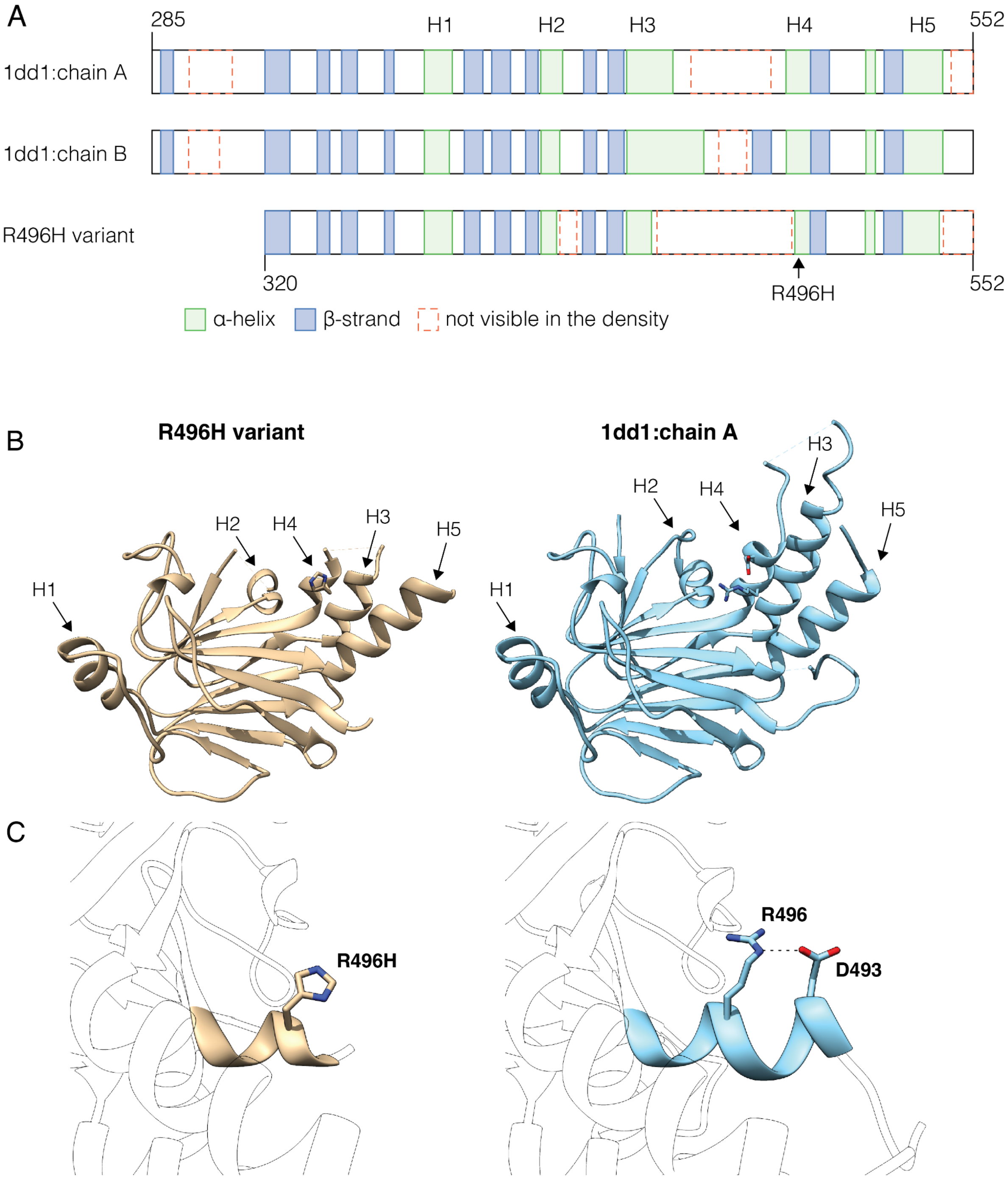


Figure 31. Crystal structure of the SMAD4 MH2 domain (residues 314-552, R496H variant).
A. Schematic representation of the secondary structural elements observed in two WT SMAD4 structures deposited in the PDB 1DD1 (top) and the structure determined here (bottom). Elements are shown as blue and green boxes. The dashed red box represents areas not visible in the electron density map. The position of the R496H variant is indicated by an arrow. B. The structure of the R496H variant is shown on the left (sand) and one of the published 1DD1 structures of the WT on the right (blue). His side chain is shown as ball-and-stick. Helices 3 and 4 in the variant are shorter than in the WT. C. A close-up view of the variation site, highlighting the differences in the length of helix 4. The contacts between R496 and D493, which stabilize the extra helical turn in the WT and that are missing in the variant, are indicated.

4.3 HTS against SMAD4 MH2 domain variants, finding small-molecule binders

As discussed in the introduction, most efforts to identify small molecules with pharmacological applications to regulate TGF-β signaling in disease have focused on identifying receptor inhibitors, or molecules that bind to the hormone and prevent it from binding to the receptor **(Akhurst, 2017; Huang *et al.*, 2021; Shi *et al.*, 2022)**. These molecules have not reached the market as treatments due to the number of side effects and complications associated with them. We set out to define a new target for the regulation of the pathway and to investigate whether SMAD4 could be an effective target for drug discovery. Our rationale for this hypothesis was twofold:

1. Considering the specificity of potential binders, we chose SMAD4 because its sequence differs more from the rest of the R-SMAD proteins, whose sequences are highly conserved.

2. SMAD4 is frequently mutated in several diseases, and different mutations determine the complexes between SMAD4 and R-SMADs.

At this stage, our search for small molecule binders was not restricted to a specific type of hit (activators, inhibitors, allosteric modulators or binders) since there are a broad range of potential applications for each type of molecule in fundamental and applied research. SMAD4 is a hub in TGF-β signaling, and interacts with various proteins (other SMADs, several activators and repressors) possibly using several binding sites. We would like to identify molecules specifically able to modulate some of these interactions in order to enhance specific aspects of TGF-β signaling -as tumor suppressor- and reduce other effects -as tumor promoters-. These hit molecules (either binders or activators/repressors) may have pharmacological applications or can be used as new research tools.

These tools are highly sought after by the research community as they would certainly open new avenues to discover novel SMAD protein binders in cells (activators or repressors) that may have been overlooked. They will also accelerate research on SMAD function and TGF-β signaling in search of new insights into the mechanisms underlying tumor development and metastasis progression. Insights that will help classify tumors at an early stage and apply tailored medicine to patients.

Perhaps, in some favorable cases, these molecules can be used as chaperones to stabilize some of these mutant SMAD4 proteins or can be combined to make bifunctional molecules to facilitate proteasome degradation. We are also interested in molecules that can decrease the thermal stability of SMAD4, since some mutations increase the stability of the protein, as those observed in the MyS, making it more resistant to degradation.

With this idea in mind, we performed two screening campaigns, with SMAD4 protein as target, and also including some of the MyS variants. In the first campaign, we screened the EU-OPENSCREEN Libraries experimentally, which contain a few more than 100 000 compounds. In the second campaign, we used a small library containing FDA or EMA approved drugs, the Prestwick Chemical Library® (PCL), as well as molecules tested in preclinical or clinical studies or approved for veterinary use. For the compounds identified in the PCL screening, if we can prove that they are useful for MyS individuals, we could identify a potentially repurposed use of some of these drugs as a medication for individuals that so far do not have treatments.

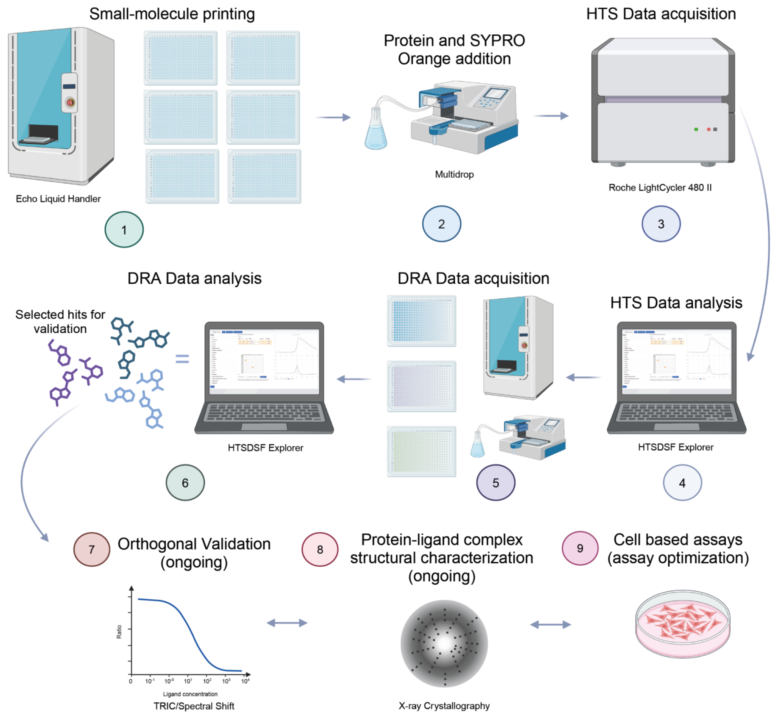


Figure 32. Pipeline for automated DSF-based high throughput screening (HTS).
(1) Compounds are printed into 384-well plates using an ECHO liquid handler (EU-OPENSCREEN, Oslo). In this step, 0.25 µL of 10 mM compound, 0.15 µL DMSO and 0.1 µL mQ water are added to 384-well plates in columns 3 to 22. Columns 1,2,23 and 24 are filled with 0.4 µL DMSO and 0.1 µL mQ water. Plates are sealed and stored at -20 ºC. (2,3) Plates were thawed in groups and filled with a solution of Sypro Orange, protein and protein buffer. We added 9.5 µL of this solution for a final well volume of 10 µL, 5X Sypro Orange and 0.1 mg/ml SMAD4. (4) Plates were run on a qPCR system and analyzed using HTSDSF Explorer. (5,6) Selected hits are validated by dose response assays (DRA). (7) Orthogonal validation using a different biophysical assay. (8) Best hits are currently being tested for crystallization (9) Cell based assays are currently being optimized to follow the effect of the compounds in a native context.

The screening was based on the application of Differential Scanning Fluorimetry (DSF), which is a fast and affordable technique to determine protein melting temperature (T_m). T_m changes can be used to indicate small molecule binding when searching for hits with potential applications in drug discovery **(Figure 32)**. These changes can be either an increase or decrease in the final T_m. Due to the large number of compounds being analyzed, the assay has been miniaturized for use in 384-well plates. The technique has gained wide acceptance as a method for the easy and rapid screening of large libraries of compounds **(Martin *et al.*, 2013; Gao, Oerlemans and Groves, 2020; Støve *et al.*, 2020)**. Screening was performed using SMAD4 WT and three variants, I500V and R496C identified in MyS and R361G reported in pancreatic and digestive cancers.

In this chapter, we present the first steps towards the generation of potential pharmacological strategies based on the use of small molecules to modulate SMAD4-dependent diseases. The conclusions of this section are that we have identified a set of compounds that interact with SMAD4 MH2 domain, some of which belong to collections of bioactive drugs that have already been approved by European and American regulatory agencies. We plan to validate whether the approved doses are also active in cell lines or models of MyS and other rare diseases, as there are no treatments for individuals with these conditions.

4.3.1 Library screening

We screened 100037 compounds in search for SMAD4 MH2 WT domain binders. We identified 462 hits (0.47% hit rate). Among the identified binders, some were already described as bioactive, but also new compounds with no activity reports. Hits were identified thanks to the low standard deviation of the references in each plate **(Figure 33A)**. To facilitate the analysis of the DSF based high-throughput screening, we developed a software to easily process large amounts of data while precisely identifying hits that may be capable of stabilizing or destabilizing the SMAD4 MH2 domain. We used HTSDSF Explorer **(Martin-Malpartida *et al.*, 2022)** interface to increase our analysis speed compared to other software that required more manual intervention. Because the software generates reports for each hit, we were able to easily identify our hits and select them for dose-response experiments. The best candidates were classified according to the apparent binding affinity, **(Table 5, Figure 33B)**. DSF also allowed us to identify the capability of each compound to stabilize or destabilize the MH2 domain of SMAD4 in the experimental conditions.

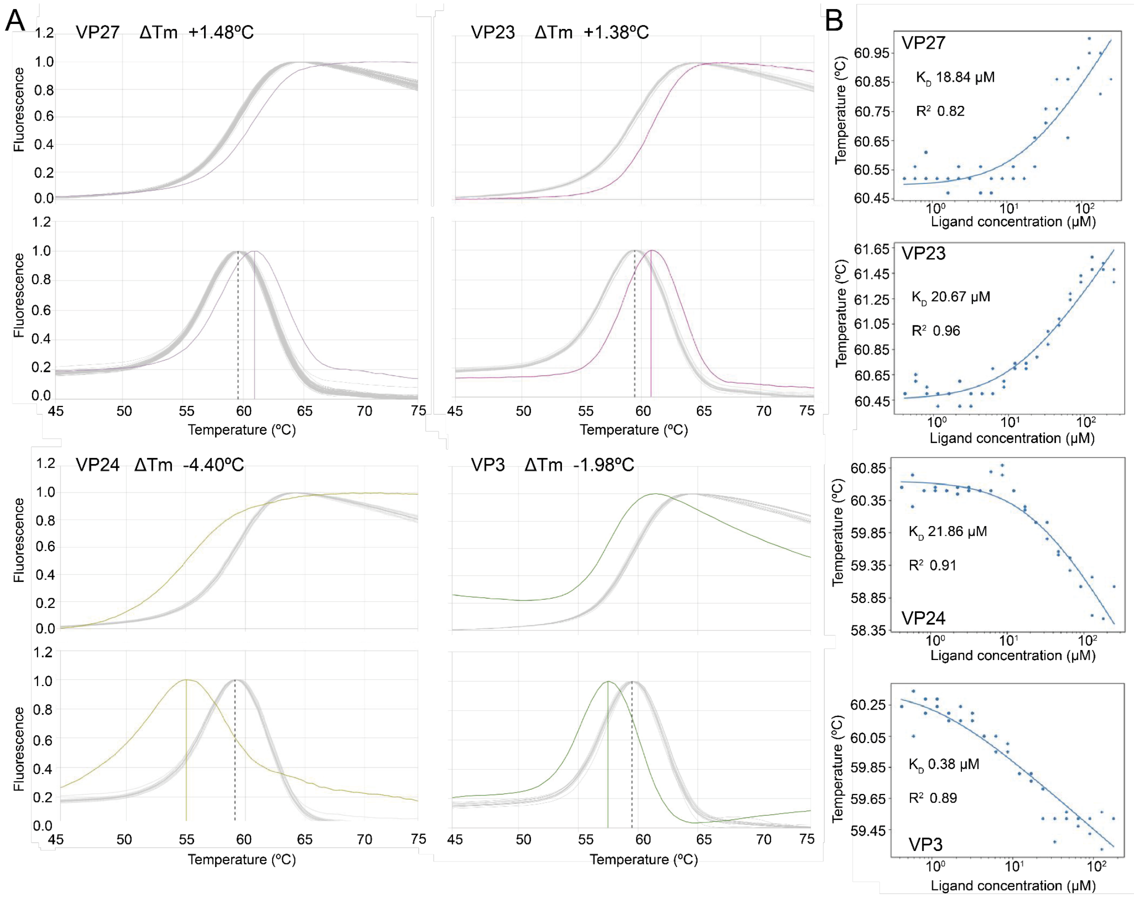


Figure 33. Protein-ligand binding experiments between SMAD4 constructs and selected T_m modulators.
A. Unfolding profile (TOP) and first derivative (BOTTOM) of SMAD4 272-552 incubated with 250 µM of VP27, VP23, VP24 or VP3 compound. References, shown in gray, have a low standard deviation which allows the selection of this ligand as a hit with the DRA curve and fitting. B. DSF dose-response curve and K_D fitting using HTSDSF Explorer. Stabilizer (VP27 and VP23) and destabilizers (VP24 and VP3) are shown.

In total, through DRA experiments we confirmed 185 compounds as validated hits (40.04% of the initial hits). 25 (13.51%) of them were considered as stabilizers and 160 (86.49%) as destabilizers **(Figure 33)**. Classifying by affinity, and if destabilizers are included in the analysis, 84 compounds were identified as high affinity binders with K_D≤100 µM **(Supplementary Table 6)** and 102 as low affinity binders with K_D>100 µM **(Supplementary Table 7)**.

4.3.2. Using EU-OPENSCREEN Library profiling for hit characterization

Available bioprofiling data for the EU- OPENSCREEN library can be used to further classify our candidates and hit selection for cell-based assays. As stated in the Material and Methods section, assays to define parameters like cell viability, luciferase reaction interference assay and ROS production have already been performed by the network and the results are available as open-access data in the ECBD website. Using this information, we flagged 25 hits that could lead to undesired cytotoxic or hepatotoxic effects in a cell viability ATP quantification assay in HepG2 cells.

4.3.3 Binding of FDA/EMA-approved drugs of the Prestwick Chemical Library (PCL)

Using the PCL with FDA and EMA approved drugs (1520 compounds), we identified several hits for SMAD4 WT and three variants. We performed the primary screening campaign following the same protocols as we did for the EU-OPENSCREEN library using DSF. In this case, however, we validated the primary hits using the Dianthus system available at the drug screening platform at the IRB Barcelona. 13 hits were validated as binders of either the WT or MyS variants, and all of them are destabilizers. Almost all validated hits were able to interact with all tested variants, with APC-44 binding with high affinity to all of them. Only APC-52, an antineoplastic agent that shows dose-dependent behavior only with R361G **(Table 5)**.

Table 5. Validated hits of Prestwick Library.

The names of compounds are anonymized. The dissociation constants (K_D; in µM) calculated through DSF are presented for each of the variants. *Low affinity* means that the value was not saturated at the compound concentration used in the assay, being the $K_D \gtrsim 125µM$.

Molecule ID	WT	R361G	R496C	I500V
APC-3	Low affinity	Low affinity	K_D=80, R^2=0.71	K_D=5.54, R^2=0.8
APC-5	Low affinity	K_D=115, R^2=0.93	K_D=26, R^2=0.9	K_D=89, R^2=0.95
APC-9*	Low affinity	Low affinity	Low affinity	Low affinity
APC-19	Low affinity	K_D=13.3, R^2=0.7	Low affinity	Low affinity
APC-20*	K_D=91, R^2=0.89	K_D=76, R^2=0.6	Low affinity	Low affinity
APC-21	K_D=7.5, R^2=0.75	K_D=42, R^2=0.6	K_D=23, R^2=0.6	K_D=20, R^2=0.9
APC-23	K_D=79, R^2=0.8	K_D=25.0, R^2=0.9	K_D=17, R^2=0.94	K_D=8.3, R^2=0.7
APC-32	Low affinity	Low affinity	Low affinity	Low affinity
APC-38	K_D=0.7, R^2=0.6	K_D=116, R^2=0.7	K_D=0.8, R^2=0.5	K_D=3.8, R^2=0.6
APC-40	K_D=5, R^2=0.8	K_D=7.3, R^2=0.9	K_D=26, R^2=0.8	K_D=18, R^2=0.9
APC-42	Low affinity	Low affinity	Low affinity	Low affinity
APC-44	K_D=3.3, R^2=0.8	K_D=0.6, R^2=0.7	K_D=0.8, R^2=0.7	K_D=0.4, R^2=0.6
APC-52		Low affinity		

We observed two compounds having greater affinity for MyS variants (APC-3 and APC-5), others for the cancer-related variant R361G (such as APC-20) or for the WT (APC-21). APC-40 binds with high affinity to the WT and also to R496C. We also noticed that APC-42 is being used as an anti-inflammatory and for the reduction of polyps in familial adenomatous polyposis, which may be of interest for the treatment of primary tumors of the gastrointestinal tract and JPS/HHT. However, this compound is bound with low affinity and may require further modification or attachment to a bio-PROTAC to increase its binding specificity and potential applicability.

From the library, we purchased 11 compounds for additional validation through TRIC and Spectral Shift. We will validate these molecules by means of biophysical and cellular assays using mouse or human cell lines. We will use cells (of both sexes), in enough quantity as to ensure that any difference observed in the experiments is statistically relevant. All selected disease mutants (Myhre syndrome and cancer) affect both men and women, indistinctly. As we mentioned in the introduction, validated hits that bind with good to medium affinity will not be discarded completely, as they could be derived as new efficient PROTAC molecules or as chemical probes as follow-up projects.

4.3.4 Structural characterization of hits binding to SMAD4 MH2 domain using Pipedream and CRIMS, EMBL Grenoble

Using the selected hits from the EU-OPENSCREEN and Prestwick Chemical libraries, we are currently performing crystal studies to describe their binding sites. We are also starting in-cell validation assays.

Regarding the X-ray crystallography, to ensure that the conditions are reproducible, we got access to the EMBL-Grenoble platform, which allows for fully automated crystal mounting, data collection, processing and calculation of initial models, thanks to their CrystalDirect technology. Data collection is performed at the beamlines of the European Synchrotron Radiation Facility (ESRF), operated by the ESRF-EMBL Joint Structural Biology Group.

As we had previous experience in crystallizing SMAD4, we reproduced these crystallization conditions in the HTX platform as a starting point. After diffraction and automated data processing, we selected 0.2 M ammonium sulfate and 26% PEG 3350 as the best condition for the project. Regarding the compounds, as they show low water solubility, they are dissolved in DMSO. Thus, before starting the project, we determined that the protein crystals were stable at up to 11% DMSO.

In a first round of HTX, we used the crystal soaking strategy **(Wienen-Schmidt *et al.*, 2021)**. In this procedure, the protein crystal is exposed to a solution of the ligand. Since the crystal has trapped water molecules, the hits can diffuse through these water-bound regions and then interact with the crystallized protein. We soaked the crystals with the first 32 hits, using a 100 mM compound stock solution, for a final concentration of 11.1 mM in the soaking solution, except for VP8 and VP26, which had low solubility in DMSO and had a concentration of 2.78 mM. We obtained three initial datasets with a real-space correlation coefficient higher than 0.7 for compounds VP12, VP21 and VP32. However, the resolution of the refined structures was only between 3.0 Å and 3.5 Å, and the ligand occupancy was low. The compounds occupied shallow cavities, often present in PPI domains. VP21 and VP32 datasets were further processed with phenix.polder using the Phenix GUI, manually selecting the region of interest, which was either the ligand or the side chains of the cavity **(Figure 34)**.

In a second round of soaking trials, we lowered the concentration of these compounds as the concentration was too high, and they precipitated in the presence of the water surrounding the crystals. By doing this, we are thus increasing the compound availability. Moreover, we also included in the HTX trial some new hits from the Prestwick Chemical Library. In parallel to soaking, we also tried co-crystallization as this method yields more reliably poses for the bound compound because it uses the mixed protein and ligand to crystallize the complex. For the co-crystallization, we used a sparse matrix screen with a constant concentration of 0.2 M ammonium sulfate and a concentration range of PEG3350 from 23% to 26% in 1% increments. Co-crystallization solutions were prepared using a Mosquito Crystal robot and manually mixed with a multichannel micropipette prior to dispensing into the CrystalDirect CD3 plate.

When we solved the structures of the protein control (in the absence of ligands), we observed the presence of ammonium sulfate bound to the protein, which could interfere with ligand binding. This led us to use a reverse soaking strategy with the objective of removing the presence of ammonium sulfate, which was done after manual harvesting, and in a 20% PEG 3350, 15% glycerol, 10% ethylene glycol, 2 mM ligand, 2% DMSO solution, at different incubation times. We have diffracted the many crystals that we have obtained. Analysis of these results is ongoing at the time of writing. For the remaining compounds, we are also validating their properties and possible strategies for using them to develop derivatives such as PROTACs. These are bifunctional molecules containing three components: the protein of interest (POI) binding moiety, a linker and the E3 ubiquitin ligase binding moiety. These derivatives may have applications in conditions such as MyS, where SMAD complexes are stabilized and accumulate.

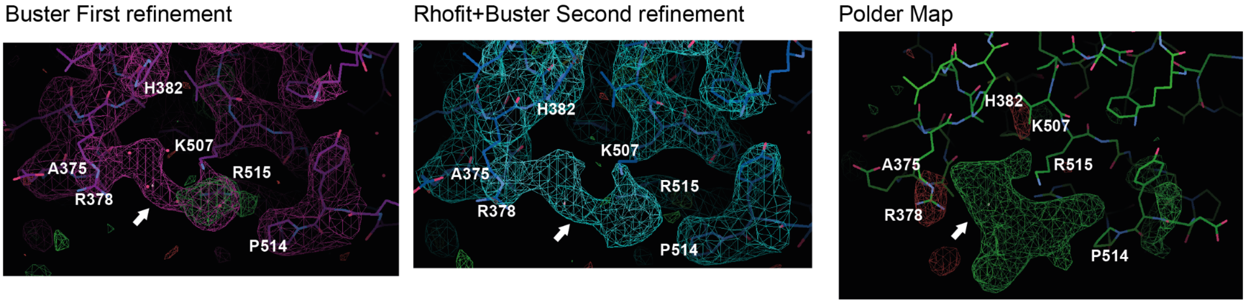


Figure 34. Optimization of VP21-bound SMAD4 electron density through Pipedream and Polder Maps (Phenix GUI).

Potential VP21-interacting SMAD4 residues are highlighted. Ligand density is indicated with a white arrow.

4.4 DNA binding properties of RREB1 ZF14-15

4.4.1 Complex Structure with the GGTCCT motif

As we mentioned in the introduction, SMAD4 not only forms quaternary structures through interactions with other SMADS. Once the complex is formed, they associate with other cofactors to modulate their function in a cellular context. One of these partners is RREB1. This protein contains several domains that directly interact with DNA and bring SMAD proteins to the proximity of Transcription starting sites and enhancers to activate or repress transcription. To investigate the DNA-binding capability of RREB1 ZFs, we used a fragment of the *SerpinE1* promoter sequence containing the GGTCCT site described in the literature as the binding site of this pair of ZFs, ZF14-15. *SerpinE1* is one of the known targets of RREB1. ZF14 and 15 are connected by a Krüppel linker, a highly conserved, 7 amino acids long sequence commonly found in ZF containing proteins. Before the structural work, we first estimated the binding affinity using EMSA assays, through quantification of DNA bound fraction, and later using ITC. We obtained a K_D of 69 nM and confirmed the 1:1 stoichiometry **(Figure 35)**.

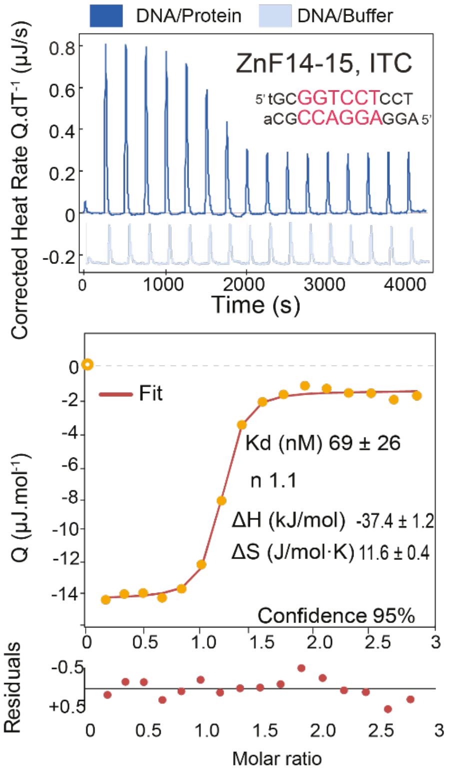


Figure 35. ITC measurement of the binding reaction between RREB1 ZF 14-15 with GGTCCT motif. Measurements were performed as stated in the Materials and Methods section.

Since the binding assays confirmed that the ZF14-15 pair could bind with good affinity to the GGTCCT sequence, we set up several crystallization experiments to study these interactions with dsDNA of different lengths and with the motif located at one site of the sequence or centered. The best diffracting crystals were obtained with a 12-mer containing the GGTCCT motif in the middle of the sequence. The structure of the complex has been refined at 1.15 Å high-resolution. The crystallographic asymmetric unit contains a copy of a protein-DNA complex in which each ZF makes specific interactions with half of the motif and the pair wraps around almost all the DNA. The schematic representation of the secondary structure elements of the 14-15 pair and the complex with DNA is shown in **Figure 36 A and B**. A summary of the data collection and refinement statistics are given in **Supplementary Table 8**. As it happens very often in protein-DNA complexes, the DNA shape is slightly distorted to accommodate the two protein helices, one from each ZF **(Figure 36A)**.

ZF14 binds to the second part of the 4-GGTCCT-9 site, through specific base contacts between Gln1580 and Arg1584 (in the α-helix) and Guanine 5 and Adenine 4 nucleotides in the complementary strand, and from Asp1581 with Cytosine 8, in the primary strand.

The orientation of the Arg1584 side chain is stabilized by hydrogen bonds from the guanidinium group to the carboxylate group of Gln1580 and by the presence of a chlorine anion, probably retained during protein purification. The chlorine is also surrounded by the guanidinium group Arg1587. Both His1585, which also coordinates Zn^{2+}, and Lys1574 (located at the second β-strand) contact the backbone DNA **(Figure 36B)**. As observed in other complexes previously described, the Thr caps the C-terminus of the ZF14 helix **(Wolfe, Nekludova and Pabo, 2000)**.

ZF15 interacts with the first part of the 4-GGTCCT-9 site. In this case, there are specific contacts between Arg1612 and Guanine 4 and His1608 with Guanine 5 and Thymine 6 bases (the latter with a suboptimal geometry) in the primary strand. In addition, due to a bend of the DNA, the protein can make abundant contacts with the backbone (phosphate groups), including interactions from Arg1602 residue located in the second β-strand as well as from and Thr1605 and from Ser1609 and His1613 residues in the α-helix itself. These contacts are schematically represented in **Figure 36B**.

Moreover, we also noticed that the residues of the Krüppel linker connecting the two fingers are well-ordered, but do not contribute to specific contacts with bases. For example, Arg1593 (we use the α isoform as the reference sequence) makes water-mediated HBs to the phosphate backbone, and the turn is facilitated by side chain stacking of Pro1594 with Tyr1595. This turn also allows for proper spacing and positioning of the next finger along the DNA. The linker also interacts with the ZnF14 helix through a salt bridge between Arg1593 and Glu1598.

There are no direct contacts to the Cytidine 7 or its complementary base, which led us to believe that this base is not important for motif recognition by the ZF14-15 pair.

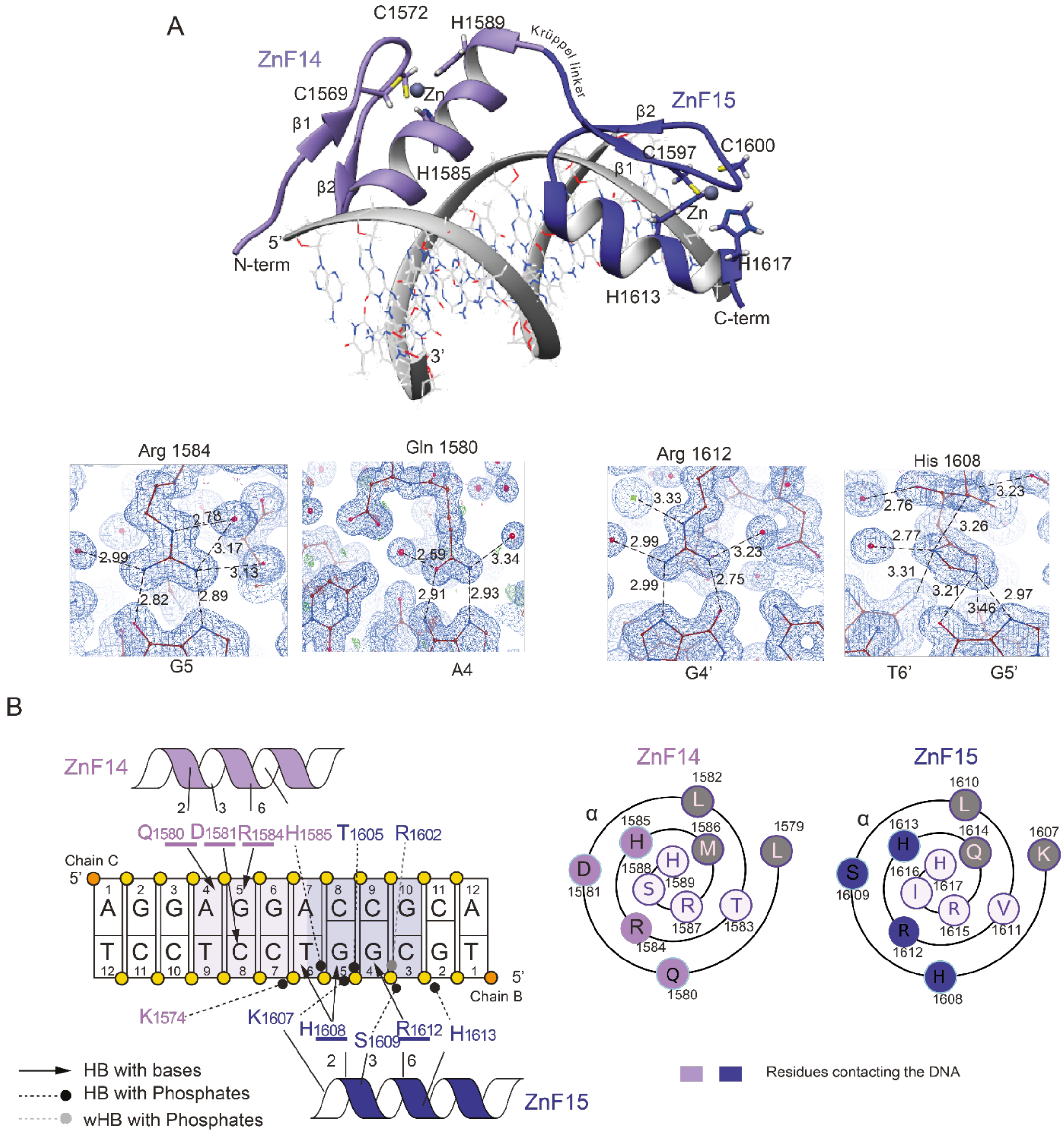


Figure 36. Complex structure of the ZF 14-15 pair bound to the GGTCCT-motif.

A. Diagram of the protein-DNA complex together with electron density maps for the key contacts of both ZFs with the DNA. B. I Contacts with bases and backbone DNA.

4.4.2 Binding to GGTCCT-like motifs

Binding to other motifs, such as the GGTCGT and GGTGCT sites also proposed in the literature, seem to require a rotation of both the His1608 and Arg1612 side chains with respect to the orientation observed in this high-resolution complex. To validate the effect of the CC to GG change, we also measured these interactions using EMSA assays. Our results revealed weak or no binding with some of the motif variants, but high affinity interaction with the motif used in the crystallization experiments (GGTCCT motif) in EMSAs **(Figure 37)**.

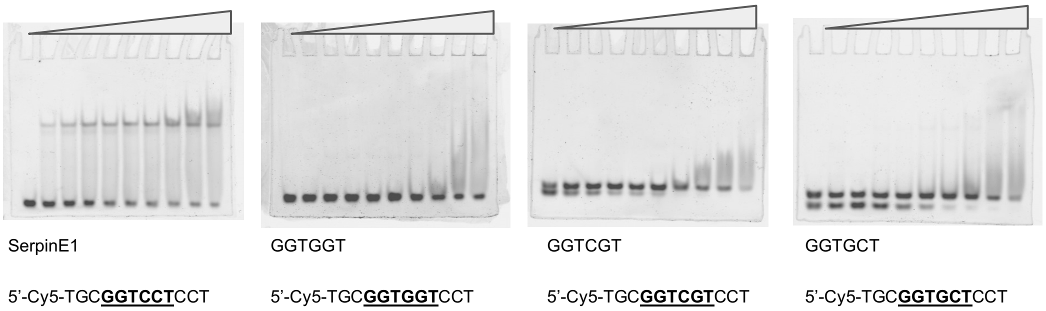


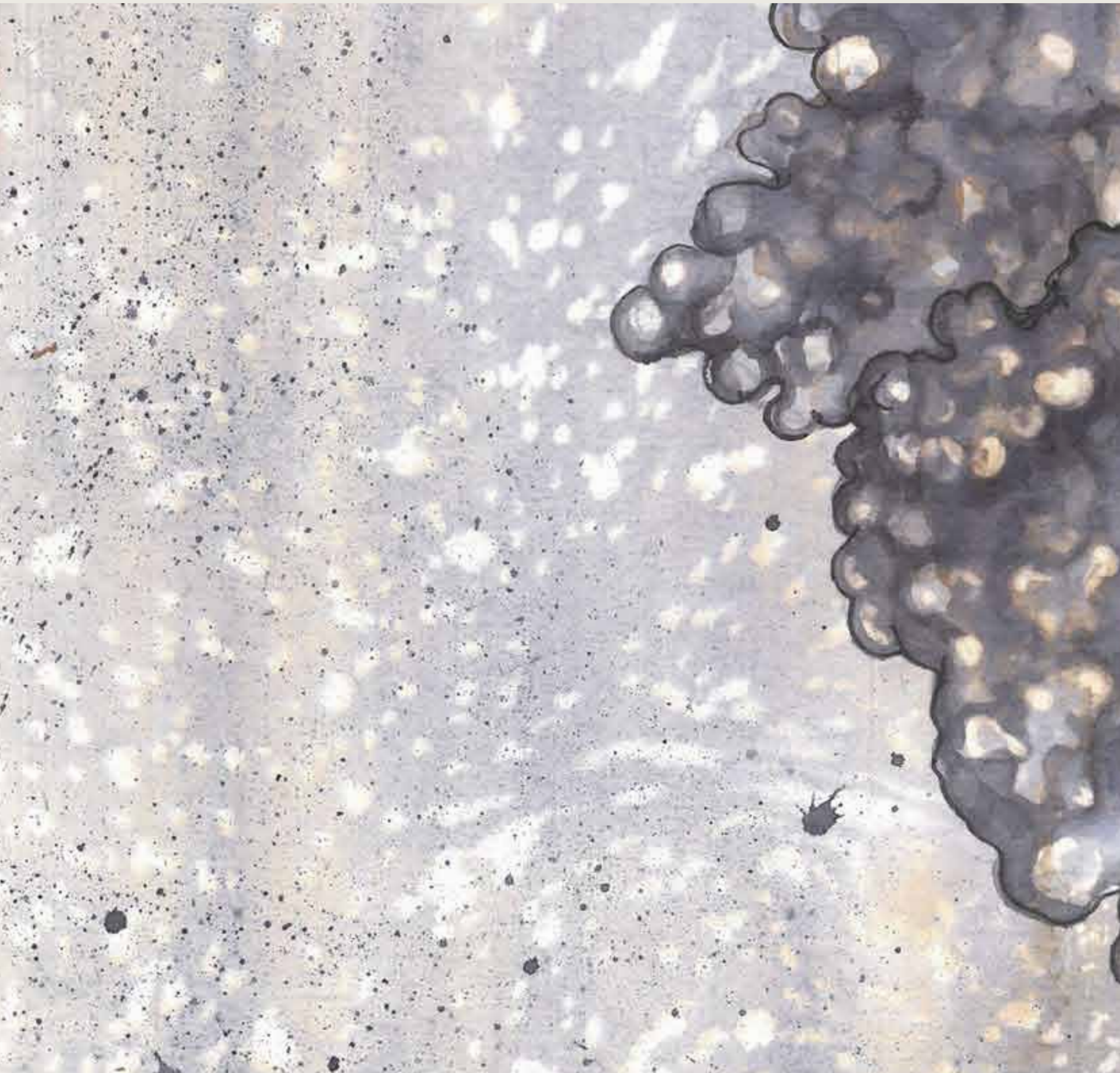
Figure 37. Comparison of binding profiles for different DNAs.

The 14-15 pair only interacts with the GGTCCT motif. Maximum concentration in the gels is 1.25 μM with a 0 μM control and 2-fold dilution factor.

Overall, the observed pattern of contacts confirms that the 14-15 pair specifically binds DNA motifs containing the GGTCCT sequences, consistent with previous descriptions in the literature.

This study is part of a broad study of the function of the protein. But I am including here the characterization of the DNA binding properties of the C-terminal part of the protein, the part that I was involved in. Part of the work was submitted for review in February of this year, and I am a co-author on it. We are also preparing a manuscript describing the structural interactions of the N- and C-terminal domains, of which I will be the first author.

See annex C for Publications related to this thesis.



DISCUSSION

5. DISCUSSION

5.1 SMAD4 variants in disease can be stratified

Since the identification of BMPs in 1965 by Marshall Urist **(Urist, 1965)** and TGFs by De Larco and Todaro in 1978 **(de Larco and Todaro, 1978)**, through the finding of the TGF-β by Harold Moses team **(Moses *et al.*, 1981)** and Michael Sporn and Anita Roberts laboratory **(Roberts *et al.*, 1981)**, and to the discovery of TGF-β receptor family **(Massague *et al.*, 1982)** and the SMADs proteins **(Sekelsky *et al.*, 1995)**, the research in this signaling pathway has seen an explosion of interest in basic and applied research.

This interest stems from the numerous biological processes that are regulated or dysregulated by this network of signaling pathways. These include essential processes such as tissue repair or embryonic development **(Wu and Hill, 2009; Marconi *et al.*, 2021; Lee and Massagué, 2022; Massagué and Sheppard, 2023)**, or their key role in numerous diseases such as fibrosis and cancer **(Puche, Saiman and Friedman, 2013; Kalluri, 2016; Marconi *et al.*, 2021; Lee and Massagué, 2022; Massagué and Sheppard, 2023)**. All these reasons have attracted the attention of numerous researchers in the fields of molecular and structural biology, genetics and medicine.

As this signaling network is highly conserved in metazoans, it has allowed the study of common and differential features in all model organisms. One of the turning points in TGF-β signaling research was the discovery that variants in SMAD4 are associated with colon and pancreatic cancer. The explosion of whole genome or exome sequencing projects has made it possible to detect numerous variants, first in many tumors and then in individuals with rare diseases. We now know that the number of patients affected by SMAD4 alterations is dependent on cancer type **(Wang *et al.*, 2021; Racu *et al.*, 2022)**. For example, in small intestine, pancreatic or colorectal cancer patients the incidence of *smad4* alterations is almost a quarter of the variants detected, but in other types like breast, melanoma or ovarian cancers, this gene is nearly unaffected (~1.5%).

Variants are annotated to reflect their position, and the relevance of a given mutation is considered based on the number of instances reported with that particular difference. During the last decades, great efforts have been made to determine protein structures, and only recently have we begun to localize them in the three-dimensional structure of proteins. In the case of SMAD proteins, we have observed that mutations often occur in clusters, in protein-protein and protein-DNA interaction regions **(Macias, Martin-Malpartida and Massagué, 2015)**. Even so, it is not trivial to predict the effects of point changes on the protein structure, let alone the effect of these differences on the protein function. We began to hypothesize that if the mutations clustered in PPI regions, maybe that could give rise to changes in the stoichiometry and stability of the quaternary structure of SMAD, which might explain some functional alterations in diseases, as suggested by other laboratories **(Caputo *et al.*, 2012; Fleming *et al.*, 2013)**. Initial findings by Benoy M Chacko, Kai Lin and co-workers **(Chacko *et al.*, 2004)** and prospective studies based on the already published structures **(Caputo *et al.*, 2012; Fleming *et al.*, 2013),** suggested a common loss-of-function effect of SMAD4 mutations in disease. Nevertheless, the described mutations at position 500 and 496 of SMAD4 and their generation of a gain-of-function effect and an increase in the amount of protein in primary cells derived from individuals with Myhre syndrome **(Caputo *et al.*, 2012, 2014)** made us though in a possible stratification of SMAD4 variants for a better assessment of its effects in patients. If so, it might be useful to categorize mutations according to their effect in SMAD association, in both tumors and rare diseases, specially if we aim at designing small molecules with pharmacological applications.

Therefore, in the first part of this thesis, we focused on studying SMAD complexes with a combination of biophysical techniques, using WT SMAD4 and disease associated variants, and SMAD3 as an example of receptor activated SMADs. We started by establishing a pipeline to study the composition of the complexes and the relative abundance of each SMAD protein in a given complex. To this end, we have tested a number of biophysical approaches until we found the optimal conditions for the study. For the SMAD proteins, the combination of DSF with mass photometry and SAXS is very suitable because it allows us to see the increase in stability when heterocomplexes are formed and quantify the different complexes.

Our findings revealed that SMAD4 variants can be stratified in different groups based on the different capability of the MH2 domain to associate with R-SMADs. For instance, MyS variants showed an increased propensity to associate with SMAD1/3 in different types of assays, and the findings were verified using different protein batches and constructs. Increase of protein affinity and decrease of dissociation rate leads to more stable complexes, which may explain the cellular phenotype previously reported **(Le Goff *et al.*, 2011)**. The prevention (or retardation) of SMAD4 targeted degradation through the proteasome could be associated with this increased complex stability and its competition with ubiquitin binding site in Lys519, which may interfere with the trimer binding interface **(Dupont *et al.*, 2009)**. We are currently determining the atomic structure of MyS variants, and we plan to continue the investigation to the complexes with R-SMADs. New questions also arise from our work as to how SMAD oligomerization equilibrium in the cell or how their specificity is translated to the regulation of genes associated with TGF-β and BMP signaling. Traditionally, the active complex in SMAD signaling pathway has been considered to be one SMAD4 molecule together with two R-SMADs, often from the same subtype. Our findings, together with the gene transcription variations upon TGF-β1 and BMP4 activation **(Alankarage *et al.*, 2022)** or basal activity of SMAD proteins **(Le Goff *et al.*, 2011; Caputo *et al.*, 2012, 2014)** reported in the bibliography, could be associated to a change of protein complex composition **(Figure 38)**. The study of these SMAD4 variants can then add information to reveal the possible specificity for certain genes of different types of SMAD complexes and even of a SMAD4 independent signaling after activation of the pathway. This could be key during embryo development, where SMAD4-independent Nodal signaling has been characterized in zebrafish embryos *in-vivo* **(Guglielmi *et al.*, 2021)**, but also in tissue homeostasis and repair. Future work could also benefit from strategies to calculate or predict the composition of SMAD complexes in primary cell lines **(Lucarelli *et al.*, 2018)**. One question that we pose to ourselves is if MyS SMAD4 variants could have an increased preference for a specific R-SMADs, or R-SMAD subgroup (TGF-β or BMP activated). We hope to contribute to this understanding in future work.

Cancer variants studied in the bibliography and the efforts of many laboratories showed the importance of changes on SMAD4 oligomerization with R-SMADs in this disease. Publication of SMAD4 complexes with SMAD2/3 revealed how many cancer mutations localized in the binding interface with the R-SMADs **(Chacko *et al.*, 2004; Fleming *et al.*, 2013)** and the effect of some of these mutations have been experimentally analyzed, showing a reduction in affinity for R-SMADs **(Chacko *et al.*, 2004)**. In our study, we selected mutations reported in *smad4* gene in cancer, shared in some cases with rare diseases such as JPS, revealing details about their effects on oligomerization. From this limited study, we proposed that variants can be classified into three major groups (four if including MyS gain-of-function variants) based on their effect in complex formation. This stratification could have a potential application in terms of patient classification and risk assessment if such differences in oligomerization end up to be clinically relevant. From this perspective, an extended study to correlate clinical data and complex formation properties of the variants is necessary, as our results do not cover numerous examples.

Our data also reveal an important factor to take into account for such analysis, and is the possible different effects produced by different amino acid changes in the same protein residue. A clear example are R496C and R496H variants in the same residue, which lead to different effects, gain-of-function and loss-of-function respectively. One of the conclusions of our analysis is that we should be very cautious before associating an effect with just a given position, since it is important to pay attention to the specific amino acid change.

As in MyS, we proposed that SMAD4 cancer and JPS associated variants can modify the SMAD complex composition in cells **(Figure 38)**.

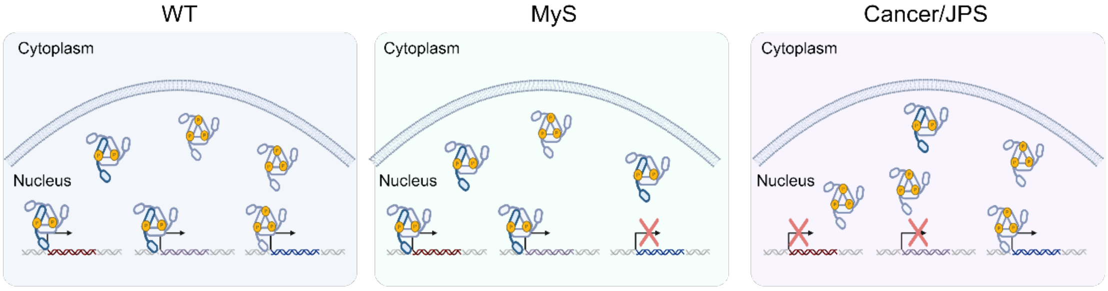


Figure 38. Proposed mechanism for gene transcription dysregulation driven by SMAD4 variants in MyS and cancer/JPS.

Assuming the existence of a SMAD4 independent TGF-β signaling, SMAD4 variants in disease could potentially affect the composition of the overall functional SMAD complexes. In the diagram, SMAD4 is shown in blue and R-SMADs in light gray.

We also determined the structure of one of the SMAD variants, R496H, a position frequently mutated in several diseases, and we are currently refining that of the same position to Cys. Given that SMAD proteins have complex architectures, we focused the structural analysis in the study of the MH2 domain, where this variant is located. Crystals of the MH2 domain show the rigid parts of the domain in high detail, while flexible regions corresponding to long loops and a long helix are not visible in the electron density. We found that, in the R496H structure, there is an increase in overall flexibility. The substitution of Arg by His has a direct influence on the secondary structure of helix 4, which is shorter than in the WT. In our group, we are now investigating whether this effect is observed in other variants and whether flexibility, together with changes in the quaternary structure of SMAD complexes, are key factors in the dysfunction of the SMAD signaling network.

The presented work can be of use to determine the best pharmacological strategy to treat patients with SMAD4 disease variants. MyS patients could benefit from SMAD4 activity inhibition strategies such as disruptors of PPIs, small molecule destabilizers or PROTACs. Meanwhile, patients with cancer and JPS SMAD4 variants could benefit from PPIs enhancers in certain stages of the disease. Drug screening campaigns as ours or the ones developed by Xiulei Mo and Haian Fu laboratories in Emory University School of Medicine **(Tang *et al.*, 2021; Ouyang *et al.*, 2024)** should take in account this type of detailed analysis for the best pharmacological strategy design.

5.2 SMAD4 small-molecule direct binders can be found and studied, supporting its druggability

The second part of the work has focused on developing an alternative approach for drugs that target the TGF-β cascade. We have moved away from the interest in blocking the receptor, both at the level of the hormone and the kinases, and have focused on SMAD4, a protein that is distributed in the cytoplasm and in the nucleus. Since variants of SMAD4 can cause both gain-of-function and loss-of-function, we were interested in identifying molecules that could act as destabilizers or stabilizers to compensate for the effect of the mutations.

In the case of individuals with rare diseases, especially in the case of Myhre syndrome, which seems associated with increased SMAD4 stability, our hypothesis was to identify compounds that can help regulate the formation and the total amount of SMADs complexes (WT and mutant), reduce their number (with compounds that prevent the formation of heterotrimers) and/or increase their degradation (for example, by promoting SMAD4 ubiquitination). We hope that our finding could be also used in diseases where an enhanced activity of SMAD4 WT protein have predominant negative effect on the tissue, like in pulmonary or liver fibrosis and advanced stages of cancers, as in Lung Adenocarcinoma, were SMAD4 may have a pivotal role to activate type 2 and type 3 EMT in association with other transcription co-factors **(Su *et al.*, 2020; Marconi *et al.*, 2021; Massagué and Sheppard, 2023)**. Our results with the identification of several destabilizers look promising in this respect. We are however aware that the identification of these hits is only the first step in early-stage drug discovery, and to develop them further as potential therapeutics, we need to keep working with experts in the next steps of drug development. In this regard, our laboratory has applied for and been granted two projects, funded by Agaur and CanServ to advance our hits to leads for potential applications as treatments for the MyS and/or for certain types of cancer respectively. We are optimistic that these combined efforts will help us to generate new lead compounds with potential pharmaceutical applications in the near future.

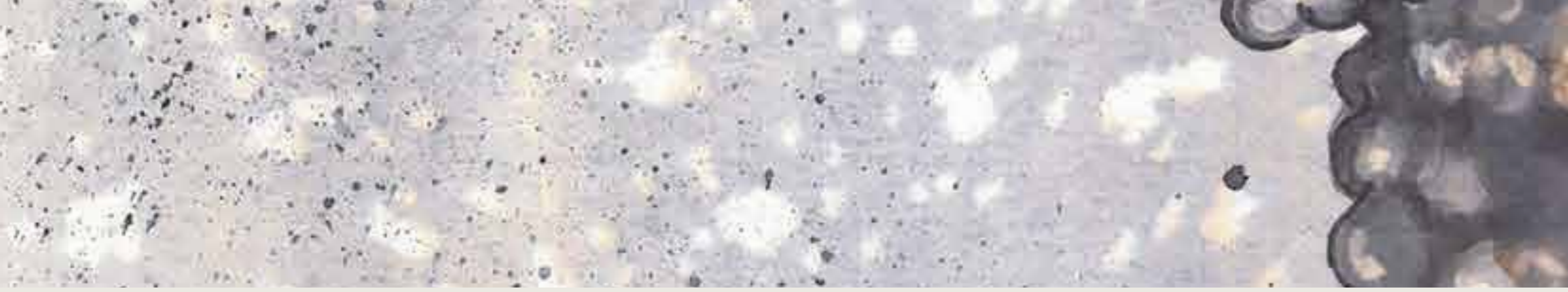
Transcription factors, as SMAD4, were considered as undruggable targets because its lack of tractable binding sites for small-molecule binding, as the ones found in kinases **(González *et al.*, 2023)**, or its structure similarity with other TFs **(Duffy and Crown, 2021; Xie *et al.*, 2023)**. We showed that with the proper technology, primary hits can be found for a TF like SMAD4. Druggability should also be put in the context of the pleiotropic nature of a gene or target of interest. In this context, SMAD4 could be a challenging target as TGF-β pathway is involved in many biological processes, and it can have a dual role in diseases even as cancer. Rare disease individuals with germ line or de novo genetic mutations could benefit the most from such advances because their gene variation is widely distributed in the organism.

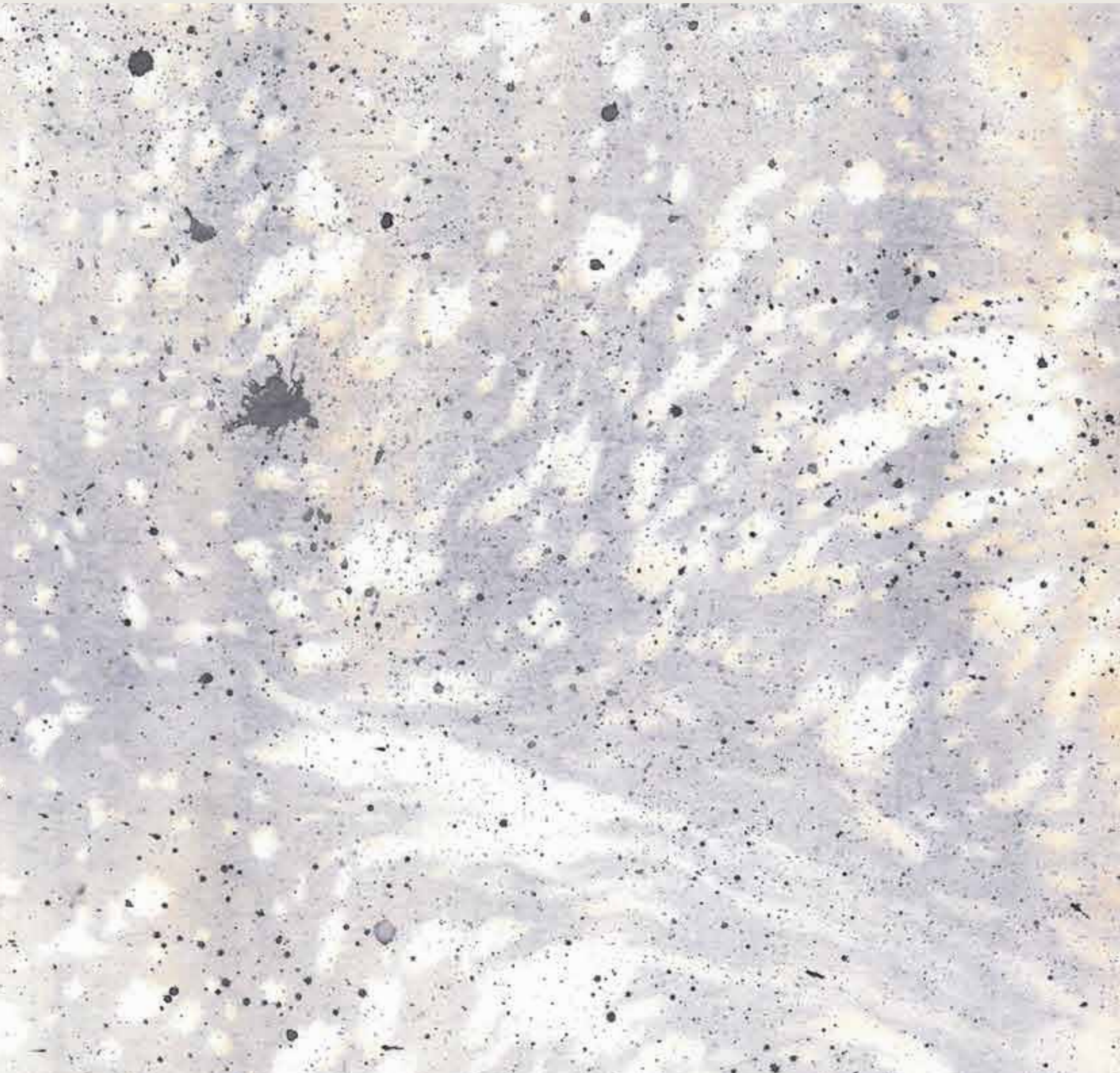
5.3 RREB1 ZFs 14-15 interact with high affinity and specificity with GGTCCT DNA motif

RREB1 is an important effector of type 2 and type 3 EMT and regulates the expression of key associated EMT TFs **(Su *et al.*, 2020)**. As a starting point to understand its functionality, we began to study its binding capability and specificity with already reported DNA motifs. ZFs 14-15 pair seem to have a strong binding with GGTCCT motif, which is also specific as proved by EMSA gels performed with variations of this DNA sequence. This interaction affinity is in the same range as other transcription factors **(Zhao *et al.*, 2018)**.

ZFs 14-15 was also reported to be a good binder of a Ras Responsive Element (RRE) identified in the calcitonin gene (5'-CCCCACCATCCCCC-3') and other genomic regions back in the 1996 (PMID: 8816445), from which a consensus sequence was proposed (5'-CCCCAAACCACCCC-3'). Paradoxically, ZFs 14-15 did not show in our hands a good binding with 5'-GGTCCT-3' or 5'-AGGACC-3' motif variations, which are highly similar with this RRE. Similarly, other labs reported that there was no-binding with RRE consensus sequence, while binding was detected with the RRE natural sequence subtracted from the calcitonin gene **(Zhang, Zhao and Edenberg, 1999)**. These differences and inconsistencies between laboratories in binding specificity of this ZF pair should be addressed in future work. We plan to further explore this ZF pair capability to interact with other DNA sequences, as starting exploring other ZFs clusters in this protein which may have other DNA binding preferences, as expected by the different zinc finger number and composition **(Najafabadi *et al.*, 2017)**.

Understanding of RREB1 DNA binding activity and the behavior of its ZFs clusters can lead to a better knowledge of how this transcription factor regulates key genes for EMT, which will also help into therapy design against these types of mechanisms. Interestingly, very little is known about RREB1 isoforms **(Nitz *et al.*, 2011)** and if they could have a more specific role in a context dependent manner in health but also in disease. If so, ZF composition may be of importance to determine the binding specificity of this protein with the DNA. In addition, RREB1 isoforms may have differences in the type of transcriptional complexes that this protein is able to form. Assessing the key elements or motifs in RREB1 in its structure responsible for such binding reactions needs to be further described. Other teams have already made advances in this field, reporting functional PXDLS motifs in RREB1 sequence responsible for complex formation with CtBP repressor complex **(Ray *et al.*, 2014)**. We hope that our future work focused on the present pipeline can help other researchers to more accurately define RREB1 function.





CONCLUSIONS

6. CONCLUSIONS

The experimental work collected in this thesis provides a biophysical, structural- and chemical biology perspective of SMAD complexes, and has advanced the process of drug discovery targeting SMAD4 to find a pharmacological solution to diseases and syndromes that so far do not have an efficient treatment. We have also begun to elucidate the structural basis of DNA recognition of the RREB1 protein, a SMAD cofactor that promotes EMT processes and drives metastatic programs.

The specific conclusions can be summarized as follows:

1. Regarding the first objective,
    a. We have established a protocol to analyze how SMAD4 variants interact with R-SMADs.
    b. SMAD4 Myhre Syndrome variants form more stable complexes with R-SMADs than the WT, whereas cancer variants display different profiles depending on the specific mutation.
    c. We have determined the structure of the R496H variant, which reveals the effects of the mutation in the fold. These effects help understand how this point mutation affects the association with R-SMADs. These effects cannot be predicted using available software, strengthening the importance of having experimental data to establish structure-function relationships.
2. Regarding the second objective,
    a. Our HTS campaign using DSF has provided the first hit binders for SMAD4.
    b. We have also identified FDA/EMA approved compounds that have been validated as hits that may have a rapid path to the clinic for the benefit of patients suffering from cancer, fibrosis and/or rare diseases.
3. Regarding the third objective,
    a. RREB1 is a multi ZF transcription factor involved in EMT processes. We have found that the ZF 14-15 pair binds the GGTCCT motif with strong affinity.
    b. We have also elucidated the key residues in the protein and the specific nucleotides that participate in the recognition.

BIBLIOGRAPHY

BIBLIOGRAPHY

Akhurst, R.J. (2017) ‘Targeting TGF-β Signaling for Therapeutic Gain’, *Cold Spring Harbor perspectives in biology*, 9(10). Available at: https://doi.org/10.1101/cshperspect.a022301.

Alankarage, D. et al. (2022) ‘Myhre syndrome is caused by dominant-negative dysregulation of SMAD4 and other co-factors’, Differentiation; research in biological diversity, 128, pp. 1–12.

Alarcón, C. *et al.* (2009) ‘Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways’, *Cell*, 139(4), pp. 757–769.

Aragón, E. *et al.* (2011) ‘A Smad action turnover switch operated by WW domain readers of a phosphoserine code’, *Genes & development*, 25(12), pp. 1275–1288.

Aragón, E. *et al.* (2012) ‘Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-β Pathways’, *Structure* , 20(10), pp. 1726–1736.

Aragón, E. *et al.* (2019) ‘Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-β signaling’, *Genes & development*, 33(21-22), pp. 1506–1524.

Attisano, L *et al.* (1993). ‘Identification of human activin and TGF beta type I receptors that form heteromeric kinase complexes with type II receptors’ *Cell,* 19;75(4), pp. 671-80.

BabuRajendran, N. *et al.* (2010) ‘Structure of Smad1 MH1/DNA complex reveals distinctive rearrangements of BMP and TGF-beta effectors’, *Nucleic acids research*, 38(10), pp. 3477–3488.

Baburajendran, N. *et al.* (2011) ‘Structural basis for the cooperative DNA recognition by Smad4 MH1 dimers’, *Nucleic acids research*, 39(18), pp. 8213–8222.

Bahceci, I. *et al.* (2017) ‘PathwayMapper: a collaborative visual web editor for cancer pathways and genomic data’, *Bioinformatics* , 33(14), pp. 2238–2240.

Bastos, M. and Velazquez-Campoy, A. (2021) ‘Isothermal titration calorimetry (ITC): a standard operating procedure (SOP)’, European biophysics journal: EBJ, 50(3-4), pp. 363–371.

Batlle, E. *et al.* (2000) ‘The transcription factor snail is a repressor of E-cadherin gene expression in epithelial tumour cells’, *Nature cell biology*, 2(2), pp. 84–89.

Brocchieri, L. and Karlin, S. (2005) ‘Protein length in eukaryotic and prokaryotic proteomes’, *Nucleic acids research*, 33(10), pp. 3390–3400.

de Bruijn, I. *et al.* (2023) ‘Analysis and Visualization of Longitudinal Genomic and Clinical Data from the AACR Project GENIE Biopharma Collaborative in cBioPortal’, *Cancer research*, 83(23), pp. 3861–3867.

Bustad, H.J. *et al.* (2013) ‘Conformational stability and activity analysis of two hydroxymethylbilane synthase mutants, K132N and V215E, with different phenotypic association with acute intermittent porphyria’, *Bioscience reports*, 33(4). Available at:

https://doi.org/10.1042/BSR20130045.

de Caestecker, M.P. *et al.* (2000) 'The Smad4 activation domain (SAD) is a proline-rich, p300-dependent transcriptional activation domain', *The Journal of biological chemistry*, 275(3), pp. 2115–2122.

Cano, A. *et al.* (2000) 'The transcription factor snail controls epithelial-mesenchymal transitions by repressing E-cadherin expression', *Nature cell biology*, 2(2), pp. 76–83.

Cao, K., Plazzer, J.-P. and Macrae, F. (2023) 'SMAD4 variants and its genotype-phenotype correlations to juvenile polyposis syndrome', *Hereditary cancer in clinical practice*, 21(1), p. 27.

Caputo, V. et al. (2012) 'A restricted spectrum of mutations in the SMAD4 tumor-suppressor gene underlies Myhre syndrome', American journal of human genetics, 90(1), pp. 161–169.

Caputo, V. *et al.* (2014) 'Novel SMAD4 mutation causing Myhre syndrome', *American journal of medical genetics. Part A*, 164A(7), pp. 1835–1840.

Cerami, E. *et al.* (2012) 'The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data', *Cancer discovery*, 2(5), pp. 401–404.

Chacko, B.M. *et al.* (2004) 'Structural basis of heteromeric smad protein assembly in TGF-beta signaling', *Molecular cell*, 15(5), pp. 813–823.

Cho, B.C. *et al.* (2020) 'Bintrafusp alfa, a bifunctional fusion protein targeting TGF-β and PD-L1, in advanced squamous cell carcinoma of the head and neck: results from a phase I cohort', *Journal for immunotherapy of cancer*, 8(2). Available at: https://doi.org/10.1136/jitc-2020-000664.

Cornaciu, I. et al. (2021) 'The Automated Crystallography Pipelines at the EMBL HTX Facility in Grenoble', Journal of visualized experiments: JoVE [Preprint], (172). Available at: https://doi.org/10.3791/62491.

Date, S. *et al.* (2004) 'Finb, a multiple zinc finger protein, represses transcription of the human angiotensinogen gene', *International journal of molecular medicine*, 13(5), pp. 637–642.

Da Vela, S. and Svergun, D.I. (2020) 'Methods, development and applications of small-angle X-ray scattering to characterize biological macromolecules in solution', Current research in structural biology, 2, pp. 164–170

David, C.J. *et al.* (2016) 'TGF-β Tumor Suppression through a Lethal EMT', *Cell*, 164(5), pp. 1015–1030.

Deng, Y.-N. *et al.* (2020) 'Transcription Factor RREB1: from Target Genes towards Biological Functions', *International journal of biological sciences*, 16(8), pp. 1463–1473.

Duffy, M.J. and Crown, J. (2021) 'Drugging "undruggable" genes for cancer treatment: Are we making progress?', International journal of cancer. Journal international du cancer, 148(1), pp. 8–17.

Dupont, S. et al. (2009) 'FAM/USP9x, a Deubiquitinating Enzyme Essential for TGFβ Signaling, Controls Smad4 Monoubiquitination', Cell, 136(1), pp. 123–135.

Feng, X.-H. and Derynck, R. (2005) 'Specificity and versatility in tgf-beta signaling through Smads', *Annual review of cell and developmental biology*, 21, pp. 659–693.

Fleming, N.I. et al. (2013) 'SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer', Cancer research, 73(2), pp. 725–735.

Fu, C. *et al.* (2021) 'Cataract-causing mutations L45P and Y46D impair the thermal stability of γC-crystallin', *Biochemical and biophysical research communications*, 539, pp. 70–76.

Fuentealba, L.C. *et al.* (2007) 'Integrating patterning signals: Wnt/GSK3 regulates the duration of the BMP/Smad1 signal', *Cell*, 131(5), pp. 980–993.

Gao, J. *et al.* (2013) 'Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal', *Science signaling*, 6(269), p. l1.

Gao, K., Oerlemans, R. and Groves, M.R. (2020) 'Theory and applications of differential scanning fluorimetry in early-stage drug discovery', *Biophysical reviews*, 12(1), pp. 85–104.

Gao, Q. *et al.* (2021) 'Knockdown of RREB1 inhibits cell proliferation via enhanced p16 expression in gastric cancer', *Cell cycle* , 20(23), pp. 2465–2475.

Gomes, T. *et al.* (2021) 'Conformational landscape of multidomain SMAD proteins', *Computational and structural biotechnology journal*, 19, pp. 5210–5224.

González, L. et al. (2023) 'Characterization of p38α autophosphorylation inhibitors that target the non-canonical activation pathway', Nature communications, 14(1), p. 3318.

Guca, E. *et al.* (2018) 'TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF-β signaling', *Nucleic acids research*, 46(17), pp. 9220–9235.

Guglielmi, L. et al. (2021) 'Smad4 controls signaling robustness and morphogenesis by differentially contributing to the Nodal and BMP pathways', Nature communications, 12(1), p. 6374.

Hellman, L.M. and Fried, M.G. (2007) 'Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions', Nature protocols, 2(8), pp. 1849–1861.

Hinck, A.P., Mueller, T.D. and Springer, T.A. (2016) 'Structural Biology and Evolution of the TGF-β Family', *Cold Spring Harbor perspectives in biology*, 8(12). Available at: https://doi.org/10.1101/cshperspect.a022103.

Hopkins, J.B., Gillilan, R.E. and Skou, S. (2017) 'BioXTAS RAW: improvements to a free open-source program for small-angle X-ray scattering data reduction and analysis', Journal of applied crystallography, 50(Pt 5), pp. 1545–1553.

Huang, C.-Y. *et al.* (2021) 'Recent progress in TGF-β inhibitors for cancer therapy', *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 134, p. 111046.

Huminiecki, L. *et al.* (2009) 'Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom', *BMC evolutionary biology*, 9, p. 28.

Hustedt, J.W. and Blizzard, D.J. (2014) 'The controversy surrounding bone morphogenetic proteins in the spine: a review of current research', *The Yale journal of biology and medicine*, 87(4), pp. 549–561.

Janda, E. *et al.* (2002) 'Ras and TGF[beta] cooperatively regulate epithelial cell plasticity and metastasis: dissection of Ras signaling pathways', *The Journal of cell biology*, 156(2), pp. 299–313.

Kalluri, R. (2016) 'The biology and function of fibroblasts in cancer', Nature reviews. Cancer, 16(9), pp. 582–598.

Kashima, R. and Hata, A. (2018) 'The role of TGF-β superfamily signaling in neurological disorders', *Acta biochimica et biophysica Sinica*, 50(1), pp. 106–120.

Kent, O.A., Fox-Talbot, K. and Halushka, M.K. (2013) 'RREB1 repressed miR-143/145 modulates KRAS signaling through downregulation of multiple targets', *Oncogene*, 32(20), pp. 2576–2585.

Konarev, P.V. et al. (2003) 'PRIMUS: a Windows PC-based system for small-angle scattering data analysis', Journal of applied crystallography, 36(5), pp. 1277–1282.

Langer, A. *et al.* (2022) 'A New Spectral Shift-Based Method to Characterize Molecular Interactions', *Assay and drug development technologies*, 20(2), pp. 83–94.

de Larco, J.E. and Todaro, G.J. (1978) 'Growth factors from murine sarcoma virus-transformed cells', Proceedings of the National Academy of Sciences of the United States of America, 75(8), pp. 4001–4005.

Lee, J.E. *et al.* (2023) 'Vactosertib, TGF-β receptor I inhibitor, augments the sensitization of the anti-cancer activity of gemcitabine in pancreatic cancer', *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie*, 162, p. 114716.

Lee, J.H. and Massagué, J. (2022) 'TGF-β in developmental and fibrogenic EMTs', Seminars in cancer biology, 86(Pt 2), pp. 136–145.

Le Goff, C. *et al.* (2011) 'Mutations at a single codon in Mad homology 2 domain of SMAD4 cause Myhre syndrome', *Nature genetics*, 44(1), pp. 85–88.

Le Goff, C., Michot, C. and Cormier-Daire, V. (2014) 'Myhre syndrome', *Clinical genetics*, 85(6), pp. 503–513.

Li, H. *et al.* (2023) 'The U2AF65/circNCAPG/RREB1 feedback loop promotes malignant phenotypes of glioma stem cells through activating the TGF-β pathway', *Cell death & disease*, 14(1), p. 23.

Liu, S., Ren, J. and Ten Dijke, P. (2021) 'Targeting TGFβ signal transduction for cancer therapy', *Signal transduction and targeted therapy*, 6(1), p. 8.

Liu, T. and Feng, X.-H. (2010) 'Regulation of TGF-beta signalling by protein phosphatases', *Biochemical Journal*, 430(2), pp. 191–198.

Lo, R.S., Wotton, D. and Massagué, J. (2001) 'Epidermal growth factor signaling via Ras controls the Smad transcriptional co-repressor TGIF', *The EMBO journal*, 20(1-2), pp. 128–136.

Lucarelli, P. *et al.* (2018) 'Resolving the Combinatorial Complexity of Smad Protein Complex Formation and Its Link to Gene Expression', *Cell systems*, 6(1), pp. 75–89.e11.

Luo, K. *et al.* (1999) 'The Ski oncoprotein interacts with the Smad proteins to repress TGFbeta signaling', *Genes & development*, 13(17), pp. 2196–2206.

Macias, M.J. *et al.* (1996) 'Structure of the WW domain of a kinase-associated protein complexed with a proline-rich peptide', *Nature*, 382(6592), pp. 646–649.

Macias, M.J. *et al.* (2000) 'Structural analysis of WW domains and design of a WW prototype', *Nature structural biology*, 7(5), pp. 375–379.

Macias, M.J., Martin-Malpartida, P. and Massagué, J. (2015) 'Structural determinants of Smad function in TGF-β signaling', *Trends in biochemical sciences*, 40(6), pp. 296–308.

Macias, M.J., Wiesner, S. and Sudol, M. (2002) 'WW and SH3 domains, two different scaffolds to recognize proline-rich ligands', *FEBS letters*, 513(1), pp. 30–37.

Marconi, G.D. et al. (2021) 'Epithelial-Mesenchymal Transition (EMT): The Type-2 EMT in Wound Healing, Tissue Regeneration and Organ Fibrosis', Cells , 10(7). Available at: https://doi.org/10.3390/cells10071587.

Martin, I. *et al.* (2013) 'Screening and evaluation of small organic molecules as ClpB inhibitors and potential antimicrobials', *Journal of medicinal chemistry*, 56(18), pp. 7177–7189.

Martin-Malpartida, P. *et al.* (2017) 'Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors', *Nature communications*, 8(1), p. 2070.

Martin-Malpartida, P. *et al.* (2022) 'HTSDSF Explorer, A Novel Tool to Analyze High-throughput DSF Screenings', *Journal of molecular biology*, 434(11), p. 167372.

Martin-Malpartida, P. *et al.* (2024) 'TPPU_DSF: A Web Application to Calculate Thermodynamic Parameters Using DSF Data', *Journal of molecular biology*, p. 168519.

Massague, J. et al. (1982) 'Affinity labeling of a transforming growth factor receptor that does not interact with epidermal growth factor', Proceedings of the National Academy of Sciences of the United States of America, 79(22), pp. 6822–6826.

Massagué, J. (2000) 'How cells read TGF-beta signals', *Nature reviews. Molecular cell biology*, 1(3), pp. 169–178.

Massagué, J. (2012) 'TGFβ signalling in context', *Nature reviews. Molecular cell biology*, 13(10), pp. 616–630.

Massagué, J., Seoane, J. and Wotton, D. (2005) 'Smad transcription factors', *Genes & development*, 19(23), pp. 2783–2810.

Massagué, J. and Sheppard, D. (2023) 'TGF-β signaling in health and disease', *Cell*, 186(19), pp. 4007–4037.

Matsuura, I. *et al.* (2004) 'Cyclin-dependent kinases regulate the antiproliferative function of Smads', *Nature*, 430(6996), pp. 226–231.

Melani, M. *et al.* (2008) 'Regulation of cell adhesion and collective cell migration by hindsight and its human homolog RREB1', *Current biology: CB*, 18(7), pp. 532–537.

Ming, L. *et al.* (2013) 'Drosophila Hindsight and mammalian RREB-1 are evolutionarily conserved DNA-binding transcriptional attenuators', *Differentiation; research in biological diversity*, 86(4-5), pp. 159–170.

Miyake, J.H., Szeto, D.P. and Stumph, W.E. (1997) 'Analysis of the structure and expression of the chicken gene encoding a homolog of the human RREB-1 transcription factor', *Gene*, 202(1-2), pp. 177–186.

Miyaki, M. and Kuroki, T. (2003) ‘Role of Smad4 (DPC4) inactivation in human cancer’, *Biochemical and biophysical research communications*, 306(4), pp. 799–804.

Miyazono, K.-I. *et al.* (2018) ‘Structural basis for receptor-regulated SMAD recognition by MAN1’, *Nucleic acids research*, 46(22), pp. 12139–12153.

Moses, H.L. et al. (1981) ‘Transforming growth factor production by chemically transformed cells’, Cancer research, 41(7), pp. 2842–2848.

Motizuki, M. *et al.* (2013) ‘Oligodendrocyte transcription factor 1 (Olig1) is a Smad cofactor involved in cell motility induced by transforming growth factor-β’, *The Journal of biological chemistry*, 288(26), pp. 18911–18922.

Najafabadi, H.S. *et al.* (2015) ‘C2H2 zinc finger proteins greatly expand the human regulatory lexicon’, *Nature biotechnology*, 33(5), pp. 555–562.

Najafabadi, H.S. et al. (2017) ‘Non-base-contacting residues enable kaleidoscopic evolution of metazoan C2H2 zinc finger DNA binding’, Genome biology, 18(1), p. 167.

Nguyen, B. *et al.* (2022) ‘Genomic characterization of metastatic patterns from prospective clinical sequencing of 25,000 patients’, *Cell*, 185(3), pp. 563–575.e11.

Nieto, M.A. (2011) ‘The ins and outs of the epithelial to mesenchymal transition in health and disease’, *Annual review of cell and developmental biology*, 27, pp. 347–376.

Nitz, M.D. *et al.* (2011) ‘RREB1 transcription factor splice variants in urologic cancer’, *The American journal of pathology*, 179(1), pp. 477–486.

Ouyang, W. et al. (2024) ‘A multiplexed time-resolved fluorescence resonance energy transfer ultrahigh-throughput screening assay for targeting the SMAD4-SMAD3-DNA complex’, Journal of molecular cell biology, 15(11). Available at: https://doi.org/10.1093/jmcb/mjad068.

Pan, D. *et al.* (2005) ‘The integral inner nuclear membrane protein MAN1 physically interacts with the R-Smad proteins to repress signaling by the transforming growth factor-{beta} superfamily of cytokines’, *The Journal of biological chemistry*, 280(16), pp. 15992–16001.

Pickup, A.T., Ming, L. and Lipshitz, H.D. (2009) ‘Hindsight modulates Delta expression during Drosophila cone cell induction’, *Development* , 136(6), pp. 975–982.

Plouhinec, J.-L., Zakin, L. and De Robertis, E.M. (2011) ‘Systems control of BMP morphogen flow in vertebrate embryos’, *Current opinion in genetics & development*, 21(6), pp. 696–703.

Pluta, R. *et al.* (2022) ‘Molecular basis for DNA recognition by the maternal pioneer transcription factor FoxH1’, *Nature communications*, 13(1), p. 7279.

Powell, H.R. (2021) ‘A beginner’s guide to X-ray data processing’, *The biochemist*, 43(3), pp. 46–50.

Puglisi, R. (2022) ‘Protein Mutations and Stability, a Link with Disease: The Case Study of Frataxin’, *Biomedicines*, 10(2). Available at: https://doi.org/10.3390/biomedicines10020425.

Puche, J.E., Saiman, Y. and Friedman, S.L. (2013) ‘Hepatic stellate cells and liver fibrosis’, Comprehensive Physiology, 3(4), pp. 1473–1492.

Racu, M.-L. et al. (2022) 'The Role of SMAD4 Inactivation in Epithelial-Mesenchymal Plasticity of Pancreatic Ductal Adenocarcinoma: The Missing Link?', Cancers, 14(4). Available at: https://doi.org/10.3390/cancers14040973.

Ray, S.K. *et al.* (2014) 'CtBP and associated LSD1 are required for transcriptional activation by NeuroD1 in gastrointestinal endocrine cells', *Molecular and cellular biology*, 34(12), pp. 2308–2317.

Roberts, A.B. et al. (1981) 'New class of transforming growth factors potentiated by epidermal growth factor: isolation from non-neoplastic tissues', Proceedings of the National Academy of Sciences of the United States of America, 78(9), pp. 5339–5343

Ruiz, L. *et al.* (2021) 'Unveiling the dimer/monomer propensities of Smad MH1-DNA complexes', *Computational and structural biotechnology journal*, 19, pp. 632–646.

Sapkota, G. *et al.* (2007) 'Balancing BMP signaling through integrated inputs into the Smad1 linker', *Molecular cell*, 25(3), pp. 441–454.

Sekelsky, J.J. et al. (1995) 'Genetic characterization and cloning of mothers against dpp, a gene required for decapentaplegic function in Drosophila melanogaster', Genetics, 139(3), pp. 1347–1358.

Shi, N. *et al.* (2022) 'Research progress on drugs targeting the TGF-β signaling pathway in fibrotic diseases', *Immunologic research*, 70(3), pp. 276–288.

Shi, Y. *et al.* (1998) 'Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling', *Cell*, 94(5), pp. 585–594.

Shi, Y. and Massagué, J. (2003) 'Mechanisms of TGF-beta signaling from cell membrane to the nucleus', *Cell*, 113(6), pp. 685–700.

Sirard, C. *et al.* (1998) 'The tumor suppressor gene Smad4/Dpc4 is required for gastrulation and later for anterior development of the mouse embryo', *Genes & development*, 12(1), pp. 107–119.

Soltermann, F. *et al.* (2020) 'Quantifying Protein-Protein Interactions by Molecular Counting with Mass Photometry', *Angewandte Chemie* , 59(27), pp. 10774–10779.

Sonn-Segev, A. *et al.* (2020) 'Quantifying the heterogeneity of macromolecular machines by mass photometry', *Nature communications*, 11(1), p. 1772.

Stetefeld, J., McKenna, S.A. and Patel, T.R. (2016) 'Dynamic light scattering: a practical guide and applications in biomedical sciences', *Biophysical reviews*, 8(4), pp. 409–427.

Støve, S.I. *et al.* (2020) 'Chapter 15 - Differential scanning fluorimetry in the screening and validation of pharmacological chaperones for soluble and membrane proteins', in A.L. Pey (ed.) *Protein Homeostasis Diseases*. Academic Press, pp. 329–341.

Su, J. *et al.* (2020) 'TGF-β orchestrates fibrogenic and developmental EMTs via the RAS effector RREB1', *Nature*, 577(7791), pp. 566–571.

Tahk, MJ. *et al.* (2023) 'Fluorescence based HTS-compatible ligand binding assays for dopamine D_3 receptors in baculovirus preparations and live cells' *Front Mol Biosci.*, 10:1119157

Tang, C. et al. (2021) 'Hypomorph mutation-directed small-molecule protein-protein interaction inducers to restore mutant SMAD4-suppressed TGF-β signaling', Cell chemical biology, 28(5), pp. 636–647.e5.

Tecalco-Cruz, A.C. *et al.* (2018) ‘Transcriptional cofactors Ski and SnoN are major regulators of the TGF-β/Smad signaling pathway in health and disease’, *Signal transduction and targeted therapy*, 3, p. 15.

Terwilliger, T.C. *et al.* (2024) ‘AlphaFold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination’, *Nature methods*, 21(1), pp. 110–116.

Thiagalingam, A. *et al.* (1996) ‘RREB-1, a novel zinc finger protein, is involved in the differentiation response to Ras in human medullary thyroid carcinomas’, *Molecular and cellular biology*, 16(10), pp. 5335–5345.

Urist, M.R. (1965) ‘Bone: formation by autoinduction’, Science, 150(3698), pp. 893–899.

Vagne-Descroix, M. *et al.* (1991) ‘Isolation and characterisation of porcine sorbin’, *European journal of biochemistry / FEBS*, 201(1), pp. 53–59.

Vincent, T. *et al.* (2009) ‘A SNAIL1-SMAD3/4 transcriptional repressor complex promotes TGF-beta mediated epithelial-mesenchymal transition’, *Nature cell biology*, 11(8), pp. 943–950.

Vonrhein, C. *et al.* (2011) ‘Data processing and analysis with the autoPROC toolbox’, *Acta crystallographica. Section D, Biological crystallography*, 67(Pt 4), pp. 293–302.

Wang, Y. et al. (2021) ‘SMAD4 mutation correlates with poor prognosis in non-small cell lung cancer’, Laboratory investigation; a journal of technical methods and pathology, 101(4), pp. 463–476.

Wang, Y.-W. *et al.* (2022) ‘Unveiling the transcriptomic landscape and the potential antagonist feedback mechanisms of TGF-β superfamily signaling module in bone and osteoporosis’, *Cell communication and signaling: CCS*, 20(1), p. 190.

Wienen-Schmidt, B. *et al.* (2021) ‘Two Methods, One Goal: Structural Differences between Cocrystallization and Crystal Soaking to Discover Ligand Binding Poses’, *ChemMedChem*, 16(1), pp. 292–300.

Wolfe, S.A., Nekludova, L. and Pabo, C.O. (2000) ‘DNA recognition by Cys2His2 zinc finger proteins’, *Annual review of biophysics and biomolecular structure*, 29, pp. 183–212.

Wotton, D. *et al.* (2001) ‘The Smad transcriptional corepressor TGIF recruits mSin3’, *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, 12(9), pp. 457–463.

Wrana, J.L. *et al.* (1994) ‘Mechanism of activation of the TGF-beta receptor’, *Nature*, 370(6488), pp. 341–347.

Wu, M.Y. and Hill, C.S. (2009) ‘Tgf-beta superfamily signaling in embryonic development and homeostasis’, Developmental cell, 16(3), pp. 329–343.

Wyatt, P.J. (1993) ‘Light scattering and the absolute characterization of macromolecules’, *Analytica chimica acta*, 272(1), pp. 1–40.

Xie, X. et al. (2023) ‘Recent advances in targeting the “undruggable” proteins: from drug discovery to clinical trials’, Signal transduction and targeted therapy, 8(1), p. 335.

Yap, T.A. *et al.* (2021) ‘First-In-Human Phase I Study of a Next-Generation, Oral, TGFβ Receptor 1 Inhibitor, LY3200882, in Patients with Advanced Cancer’, *Clinical cancer research: an official journal of the American Association for Cancer Research*, pp. 6666–6676.

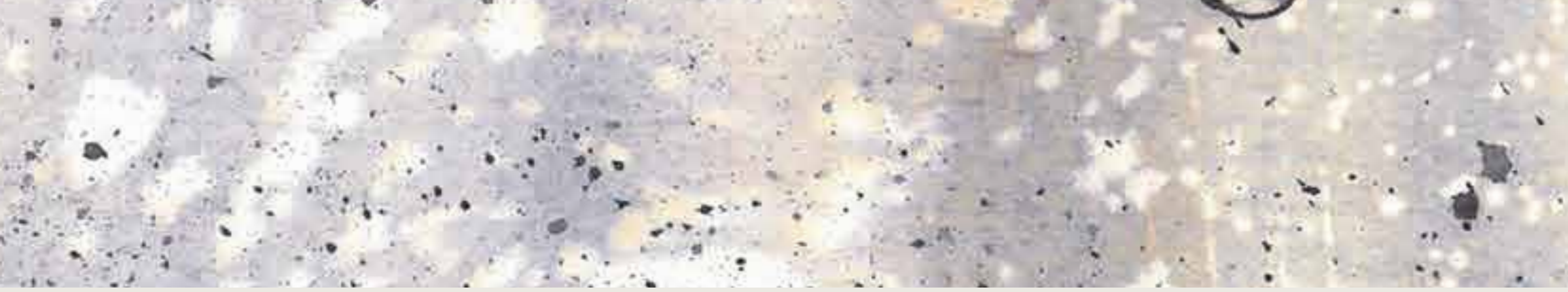
Yip, M.L., Lamka, M.L. and Lipshitz, H.D. (1997) ‘Control of germ-band retraction in Drosophila by the zinc-finger protein HINDSIGHT’, *Development* , 124(11), pp. 2129–2141.

Young, G. *et al.* (2018) ‘Quantitative mass imaging of single biological macromolecules’, *Science*, 360(6387), pp. 423–427.

Zamarioli, A. *et al.* (2022) ‘Systemic effects of BMP2 treatment of fractures on non-injured skeletal sites during spaceflight’, *Frontiers in endocrinology*, 13, p. 910901.

Zhang, L., Zhao, J. and Edenberg, H.J. (1999) ‘A human Raf-responsive zinc-finger protein that binds to divergent sequences’, *Nucleic acids research*, 27(14), pp. 2947–2956.

Zhao, Y. et al. (2018) ‘The 11th C2H2 zinc finger and an adjacent C-terminal arm are responsible for TZAP recognition of telomeric DNA’, Cell research, 28(1), pp. 130–134.



ANNEXES

Annex A. Supplementary Tables

Supplementary Table 1.

Significance of changes in ΔT_m between variants in nanoDSF experiments. The ΔT_m of each replicate is calculated with respect the average T_m value of the WT. A Welch's t-test is used to determine the significance of the means differences. MyS variants are compared with the WT. Additionally R496C was compared to the I500V variant (*).

Construct	P value	Significantly Different?	Welch-corrected t, df	Difference between means	R^2
I500V	<.001	Yes	t=5.200, df=8.018	-0.171 ± 0.032	0.771
R496C	<.001	Yes	t=57.28, df=8.646	-2.002 ± 0.035	0.997
R496C*	<.001	Yes	t=58.90, df=9.780	1.832 ± 0.031	0.997

Supplementary Table 2.

Significance of changes in ΔT_m between variants in DSF experiments. The ΔT_m of each replicate is calculated with respect the average T_m value of the WT. A Welch's t-test is used to determine the significance of the means differences. MyS variants are compared with the WT.

Construct	P value	Significantly Different?	Welch-corrected t, df	Difference between means	R^2
I500V	.021	Yes	t=4.836, df=2.703	-0.536 ± 0.111	0.896
I500T	.002	Yes	t=4.770, df=7.103	-0.260 ± 0.055	0.762
I500M	.005	Yes	t=7.833, df=2.890	-0.790 ± 0.101	0.955
R496C	<.001	Yes	t=16.18, df=7.528	-2.011 ± 0.124	0.972

Supplementary Table 3.

Significance of changes in ΔT_m between variants in DSF experiments in presence of different ratios of SMAD4 MH2 domains. The ΔT_m of each replicate is calculated with respect the average T_m value of the WT. A Welch's t-test is used to determine the significance of the means differences. MyS variants are compared with the WT.

SMAD4 R496C + SMAD3 2:1

Unpaired t test with Welch's correction	
P value	.047
P value summary	*
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=4.275, df=2.088
Difference between means (C - A) ± SEM	0.8300 ± 0.1941
95% confidence interval	0.02752 to 1.632
R squared (eta squared)	0.8975

SMAD4 R496C + SMAD3 1:1

Unpaired t test with Welch's correction	
P value	.098
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=2.705, f=2.310
Difference between means (C - A) ± SEM	0.6600 ± 0.2440
95% confidence interval	-0.2659 to 1.586
R squared (eta squared)	0.7601

SMAD4 R496C + SMAD3 1:2

Unpaired t test with Welch's correction	
P value	.162
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=2.151, df=2.032
Difference between means (C - A) ± SEM	0.6033 ± 0.2805
95% confidence interval	-0.5853 to 1.792
R squared (eta squared)	0.6948

SMAD4 R496C + SMAD3 1:4

Unpaired t test with Welch's correction	
P value	.811
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.2626, df=2.777
Difference between means (C - A) ± SEM	-0.01333 ± 0.05077
95% confidence interval	-0.1825 to 0.1559
R squared (eta squared)	0.02424

SMAD4 R496C + SMAD3 1:8

Unpaired t test with Welch's correction	
P value	.676
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.4545, df=3.476
Difference between means (C - A) ± SEM	0.01667 ± 0.03667
95% confidence interval	-0.09149 to 0.1248
R squared (eta squared)	0.05611

SMAD4 R496C + SMAD3 1:16

Unpaired t test with Welch's correction	
P value	.563
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.6701, df=2.340
Difference between means (C - A) ± SEM	-0.1033 ± 0.1542
95% confidence interval	-0.6824 to 0.4757
R squared (eta squared)	0.1610

SMAD4 R496C + SMAD3 1:32

Unpaired t test with Welch's correction	
P value	.624
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.5670, df=2.178
Difference between means (C - A) ± SEM	-0.06333 ± 0.1117
95% confidence interval	-0.5082 to 0.3815
R squared (eta squared)	0.1286

SMAD4 R361G + SMAD3 2:1

Unpaired t test with Welch's correction	
P value	0.0003
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=36.43, df=2.334
Difference between means (D - A) ± SEM	-3.723 ± 0.1022
95% confidence interval	-4.108 to -3.339
R squared (eta squared)	0.9982

SMAD4 R361G + SMAD3 1:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=25.76, df=3.813
Difference between means (D - A) ± SEM	-2.707 ± 0.1051
95% confidence interval	-3.004 to -2.409
R squared (eta squared)	0.9943

SMAD4 R361G + SMAD3 1:2

Unpaired t test with Welch's correction	
P value	.002
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=15.59, df=2.384
Difference between means (D - A) ± SEM	-1.320 ± 0.08466
95% confidence interval	-1.633 to -1.007
R squared (eta squared)	0. 9903

SMAD4 R361G + SMAD3 1:4

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=11.63, df=3.777
Difference between means (D - A) ± SEM	-0.3933 ± 0.03383
95% confidence interval	-0.4895 to -0.2972
R squared (eta squared)	0.9728

SMAD4 R361G + SMAD3 1:8

Unpaired t test with Welch's correction	
P value	.147
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=1.828, df=3.723
Difference between means (D - A) ± SEM	-0.07000 ± 0.03830
95% confidence interval	-0.1795 to 0.03952
R squared (eta squared)	0.4729

SMAD4 R361G + SMAD3 1:16

Unpaired t test with Welch's correction	
P value	.763
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.3276, df=3.325
Difference between means (D - A) ± SEM	-0.01667 ± 0.05088
95% confidence interval	-0.1700 to 0.1367
R squared (eta squared)	0.03126

SMAD4 R361G + SMAD3 1:32

Unpaired t test with Welch's correction	
P value	.712
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.4096, df=2.758
Difference between means (D - A) ± SEM	0.02333 ± 0.05696
95% confidence interval	-0.1673 to 0.2140
R squared (eta squared)	0.05736

SMAD4 I500V + SMAD3 2:1

Unpaired t test with Welch's correction	
P value	.001
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=15.71, df=2.489
Difference between means (B - A) ± SEM	1.347 ± 0.08570
95% confidence interval	1.039 to 1.654
R squared (eta squared)	0.9900

SMAD4 I500V + SMAD3 1:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=16.04, df=3.108
Difference between means (B - A) ± SEM	1.200 ± 0.07483
95% confidence interval	0.9665 to 1.434
R squared (eta squared)	0.9881

SMAD4 I500V + SMAD3 1:2

Unpaired t test with Welch's correction	
P value	.001
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=14.16, df=2.731
Difference between means (B - A) ± SEM	0.8933 ± 0.06307
95% confidence interval	0.6810 to 1.106
R squared (eta squared)	0.9866

SMAD4 I500V + SMAD3 1:4

Unpaired t test with Welch's correction	
P value	.548
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.6614, df=3.662
Difference between means (B - A) ± SEM	-0.02333 ± 0.03528
95% confidence interval	-0.1249 to 0.07827
R squared (eta squared)	0.1067

SMAD4 I500V + SMAD3 1:8

Unpaired t test with Welch's correction	
P value	.970
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.04152, df=2.659
Difference between means (B - A) ± SEM	0.003333 ± 0.08028
95% confidence interval	-0.2718 to 0.2784
R squared (eta squared)	0.0006481

SMAD4 I500V + SMAD3 1:16

Unpaired t test with Welch's correction	
P value	.395
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=1.012, df=2.641
Difference between means (B - A) ± SEM	-0.1167 ± 0.1153
95% confidence interval	-0.5134 to 0.2801
R squared (eta squared)	0.2795

SMAD4 I500V + SMAD3 1:32

Unpaired t test with Welch's correction	
P value	.977
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.03196, df=2.206
Difference between means (B - A) ± SEM	-0.003333 ± 0.1043
95% confidence interval	-0.4143 to 0.4076
R squared (eta squared)	0.0004629

SMAD4 R496C + SMAD1 2:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=8.884, df=3.935
Difference between means (C - A) ± SEM	1.167 ± 0.1313
95% confidence interval	0.7997 to 1.534
R squared (eta squared)	0.9525

SMAD4 R496C + SMAD1 1:1

Unpaired t test with Welch's correction	
P value	.003
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=15.16, df=2.189
Difference between means (C - A) ± SEM	1.090 ± 0.07188
95% confidence interval	0.8049 to 1.375
R squared (eta squared)	0.9906

SMAD4 R496C + SMAD1 1:2

Unpaired t test with Welch's correction	
P value	.001
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=9.135, df=3.785
Difference between means (C - A) ± SEM	0.8433 ± 0.09232
95% confidence interval	0.5812 to 1.106
R squared (eta squared)	0.9566

SMAD4 R496C + SMAD1 1:4

Unpaired t test with Welch's correction	
P value	.186
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=1.828, df=2.424
Difference between means (C - A) ± SEM	0.4000 ± 0.2188
95% confidence interval	-0.4000 to 1.200
R squared (eta squared)	0.5797

SMAD4 R496C + SMAD1 1:8

Unpaired t test with Welch's correction	
P value	.693
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.4482, df=2.294
Difference between means (C - A) ± SEM	0.07000 ± 0.1562
95% confidence interval	-0.5258 to 0.6658
R squared (eta squared)	0.08053

SMAD4 R496C + SMAD1 1:16

Unpaired t test with Welch's correction	
P value	.589
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.5913, df=3.621
Difference between means (C - A) ± SEM	0.06667 ± 0.1127
95% confidence interval	-0.2597 to 0.3930
R squared (eta squared)	0.08805

SMAD4 R496C + SMAD1 1:32

Unpaired t test with Welch's correction	
P value	.097
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=2.488, df=2.735
Difference between means (C - A) ± SEM	-0.1900 ± 0.07638
95% confidence interval	-0.4469 to 0.06693
R squared (eta squared)	0.6935

SMAD4 R361G + SMAD1 2:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=19.50, df=3.815
Difference between means (D - A) ± SEM	-2.163 ± 0.1110
95% confidence interval	-2.477 to -1.849
R squared (eta squared)	0.9901

SMAD4 R361G + SMAD1 1:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=18.87, df=3.587
Difference between means (D - A) ± SEM	-1.620 ± 0.08583
95% confidence interval	-1.870 to -1.370
R squared (eta squared)	0.9900

SMAD4 R361G + SMAD1 1:2

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=12.85, df=3.952
Difference between means (D - A) ± SEM	-1.097 ± 0.08537
95% confidence interval	-1.335 to -0.8585
R squared (eta squared)	0.9766

SMAD4 R361G + SMAD1 1:4

Unpaired t test with Welch's correction	
P value	.002
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=7.383, df=3.840
Difference between means (D - A) ± SEM	-0.7967 ± 0.1079
95% confidence interval	-1.101 to -0.4921
R squared (eta squared)	0.9342

SMAD4 R361G + SMAD1 1:8

Unpaired t test with Welch's correction	
P value	0.0211
P value summary	*
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=4.763, df=2.762
Difference between means (D - A) ± SEM	-0.4800 ± 0.1008
95% confidence interval	-0.8169 to -0.1431
R squared (eta squared)	0.8915

SMAD4 R361G + SMAD1 1:16

Unpaired t test with Welch's correction	
P value	.097
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=2.301, df=3.291
Difference between means (D - A) ± SEM	-0.2467 ± 0.1072
95% confidence interval	-0.5713 to 0.07801
R squared (eta squared)	0.6168

SMAD4 R361G + SMAD1 1:32

Unpaired t test with Welch's correction	
P value	.103
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=2.847, df=2.027
Difference between means (D - A) ± SEM	-0.2000 ± 0.07024
95% confidence interval	-0.4984 to 0.09835
R squared (eta squared)	0.8000

SMAD4 I500V + SMAD1 2:1

Unpaired t test with Welch's correction	
P value	<.001
P value summary	***
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=15.21, df=3.938
Difference between means (B - A) ± SEM	1.757 ± 0.1155
95% confidence interval	1.434 to 2.079
R squared (eta squared)	0.9833

SMAD4 I500V + SMAD1 1:1

Unpaired t test with Welch's correction	
P value	.002
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=9.114, df=3.248
Difference between means (B - A) ± SEM	1.257 ± 0.1379
95% confidence interval	0.8363 to 1.677
R squared (eta squared)	0.9624

SMAD4 I500V + SMAD1 1:2

Unpaired t test with Welch's correction	
P value	.003
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=7.499, df=3.308
Difference between means (B - A) ± SEM	0.8200 ± 0.1093
95% confidence interval	0.4897 to 1.150
R squared (eta squared)	0.9444

SMAD4 I500V + SMAD1 1:4

Unpaired t test with Welch's correction	
P value	.007
P value summary	**
Significantly different (P < 0.05)?	Yes
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=5.347, df=3.769
Difference between means (B - A) ± SEM	0.5933 ± 0.1110
95% confidence interval	0.2777 to 0.9090
R squared (eta squared)	0.8835

SMAD4 I500V + SMAD1 1:8

Unpaired t test with Welch's correction	
P value	.490
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.8275, df=2.153
Difference between means (B - A) ± SEM	0.1767 ± 0.2135
95% confidence interval	-0.6822 to 1.036
R squared (eta squared)	0.2413

SMAD4 I500V + SMAD1 1:16

Unpaired t test with Welch's correction	
P value	.589
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.5872, df=3.999
Difference between means (B - A) ± SEM	0.07667 ± 0.1306
95% confidence interval	-0.2858 to 0.4392
R squared (eta squared)	0.07938

SMAD4 I500V + SMAD1 1:32

Unpaired t test with Welch's correction	
P value	.894
P value summary	ns
Significantly different (P < 0.05)?	No
One- or two-tailed P value?	Two-tailed
Welch-corrected t, df	t=0.1446, df=3.231
Difference between means (B - A) ± SEM	0.02000 ± 0.1383
95% confidence interval	-0.4029 to 0.4429
R squared (eta squared)	0.006429

Supplementary Table 4.

ITC-derived binding and thermodynamic parameters of SMAD3 and SMAD4 WT, MyS variants and A406T.

Construct	K_D (M)	CI 95%	n	CI 95%
WT	5.28e-07	3.75E-07	0.55	0.02
I500V	1.72e-07	1.42E-07	0.44	0.02
I500T	1.44e-07	9.04E-08	0.52	0.02
I500M	1.06e-07	9.80E-08	0.49	0.02
A406T	2.61e-06	1.11E-06	0.53	0.03

Construct	ΔH (kJ/mol)	CI 95%	-TΔS (kJ/mol)	ΔG (kJ/mol)	ΔS (J/mol·K)
WT	-6.98	0.65	-30.29	-37.28	97.67
I500V	-19.55	1.97	-20.61	-40.16	66.44
I500T	-21.91	1.60	-18.73	-40.63	60.38
I500M	-17.68	1.65	-23.73	-41.41	76.52
A406T	-11.15	1.15	-22.01	-33.16	70.97

Supplementary Table 5.

Significance of changes in ΔT_m between variants in DSF experiments. The ΔT_m of each replicate is calculated with respect the average T_m value of the WT. A Welch's t-test to determine the significance of the means differences. MyS variants are compared with the WT.

Construct	P value	Significantly Different?	Welch-corrected t, df	Difference between means	R^2
R496C	<.001	Yes	t=16.18, df=7.528	-2.011 ± 0.124	0.972
D351G	.002	Yes	t=4.416, df=8.990	0.225 ± 0.051	0.684
P356L	<.001	Yes	t=4.776, df=9.090	-0.431 ± 0.090	0.715
R361G	<.001	Yes	t=82,21, df=6,000	-3.460 ± 0.042	0.999
G386D	<.001	Yes	t=136.7, df=4.364	-13.54 ± 0.099	0.999
A406T	<.001	Yes	t=16,72, df=7,000	-2.530 ± 0.1513	0.976
K428T	.018	Yes	t=2.902, df=8.999	-0.150 ± 0.052	0.483
R496H	.002	Yes	t=11.52, df=2.743	-1.250 ± 0.109	0.980
R515T	<.001	Yes	t=27.71, df=7.123	1.790 ± 0.065	0.991

Supplementary Table 6. High affinity binders.

Hits validated in DRA experiments by DSF with K_D lower than 100 µM. Stabilizers and destabilizers are indicated with the + and - symbol, respectively. Positive toxicity values performed in HepG2 cells (EU- OPENSCREEN data) are shown.

		T_m Shift Model					T_m Shift Model		
Molecule ID	**Effect**	**K_D (µM)**	**R^2**	**Toxic**	**Molecule ID**	**Effect**	**K_D (µM)**	**R^2**	**Toxic**
VP1	+	0.53	0.7		**M1**	-	1.48	0.85	
VP2	+	0.7	0.74	✓	**N1**	-	42.15	0.885	
VP3	-	0.38	0.89		**O1**	-	22.14	0.829	
VP4	+	0.7	0.74		**P1**	-	66.06	0.765	
VP5	-	0.804	0.854		**Q1**	-	69.02	0.948	
VP6	-	0.648	0.863	✓	**R1**	-	17.76	0.876	
VP7	-	0.376	0.711		**S1**	-	13.29	0.889	
VP8	-	0.376	0.8		**T1**	-	71.29	0.86	
VP9	-	0.376	0.72		**U1**	-	24.99	0.907	✓
VP10	-	0.375	0.85		**V1**	-	30.69	0.922	
VP11	-	0.675	0.835		**W1**	-	6.51	0.826	
VP12	-	0.376	0.7		**Y1**	-	17.27	0.916	
VP13	-	0.798	0.831		**Z1**	-	57.25	0.912	
VP14	-	0.376	0.728		**A2**	-	98.78	0.835	
VP15	-	0.376	0.807		**B2**	-	26.94	0.877	✓
VP16	-	0.34	0.71		**C2**	-	95.73	0.922	
VP17	-	2.98	0.792		**D2**	-	7.34	0.832	
VP18	-	2.1	0.748		**E2**	-	3.73	0.869	✓
VP19	-	1.03	0.899		**F2**	-	45.36	0.839	
VP20	-	1.12	0.754		**G2**	-	80.06	0.794	
VP21	-	1.01	0.751		**H2**	-	63.15	0.776	
VP22	+	1.31	0.81		**I2**	-	1.75	0.7	✓
VP23	+	20.67	0.96		**J2**	-	73.72	0.796	
VP24	-	21.86	0.91		**K2**	-	7.38	0.747	
VP25	+	0.38	0.85	✓	**L2**	-	1.36	0.712	
VP26	+	24.72	0.94		**M2**	-	10	0.843	
VP27	+	18.84	0.82		**N2**	-	12.32	0.941	
VP28	+	7.48	0.725		**O2**	-	32.34	0.882	
VP29	-	5.94	0.889		**P2**	-	12.48	0.941	
VP30	-	2.09	0.813		**Q2**	-	76.69	0.839	
VP31	-	10.88	0.884		**R2**	-	9.49	0.846	
VP32	-	2.23	0.812		**S2**	-	4.51	0.798	
A1	+	1.23	0.75		**T2**	-	5.17	0.775	
B1	+	21.01	0.77	✓	**U2**	-	2.51	0.917	
C1	+	18.97	0.73		**V2**	+	98.46	0.856	
D1	+	9.11	0.7		**W2**	-	22.81	0.926	
E1	+	45.91	0.75		**Y2**	-	37.82	0.714	✓
F1	+	4.35	0.53		**Z2**	-	23.2	0.885	

		T_m Shift Model					**T_m Shift Model**		
Molecule ID	**Effect**	**K_D (μM)**	**R^2**	**Toxic**	**Molecule ID**	**Effect**	**K_D (μM)**	**R^2**	**Toxic**
G1	-	88.17	0.99		**A3**	-	35.9	0.989	
H1	+	67	0.85		**B3**	-	6.53	0.849	✓
I1	-	52.94	0.98		**C3**	-	0.503	0.911	✓
J1	-	33.22	0.97						
K1	-	40.5	0.96						
L1	-	86.76	0.939						

Supplementary Table 7. Low affinity binders.

Hits validated in DRA experiments by DSF with K_D higher than 100 μM. Stabilizers and destabilizers are indicated with the + and - symbol, respectively.

		T_m Shift Model					T_m Shift Model		
Molecule ID	**Effect**	**K_D(μM)**	**R^2**	**Toxic**	**Molecule ID**	**Effect**	**K_D(μM)**	**R^2**	**Toxic**
D3	-	>250	0.86		**E5**	-	>250	0.826	
E3	-	>250	0.82		**F5**	-	>250	0.743	
F3	-	>250	0.83		**G5**	-	>250	0.942	
G3	-	>250	0.52	✓	**H5**	-	>250	0.804	
H3	-	>250	0.85		**I5**	-	>250	0.929	
I3	-	>250	0.58		**J5**	-	v250	0.786	
J3	-	>250	0.7	✓	**K5**	+	169.32	0.848	
K3	-	>250	0.9	✓	**L5**	-	>250	0.947	
L3	-	>250	0.9	✓	**M5**	-	>250	0.838	
M3	-	>250	0.96	✓	**N5**	-	211.27	0.939	
N3	-	>250	0.86		**O5**	-	>250	0.922	
O3	-	>250	0.89		**P5**	-	>250	0.935	
P3	-	>250	0.95	✓	**Q5**	-	>250	0.789	
Q3	-	>250	0.94	✓	**R5**	-	>250	0.819	
R3	-	>250	0.66		**S5**	-	153.12	0.89	
S3	-	>250	0.99		**T5**	-	>250	0.899	
T3	+	>250	0.76		**U5**	-	112.11	0.909	
U3	-	>250	0.89		**V5**	-	>250	0.969	
V3	-	>250	0.64	✓	**W5**	-	>250	0.833	
W3	-	>250	0.94		**Y5**	-	>250	0.829	
Y3	-	>250	0.93		**Z5**	-	>250	0.91	
Z3	-	>250	0.84	✓	**A6**	-	>250	0.78	
A4	-	102.77	0.94		**B6**	-	>250	0.824	
B4	-	>250	0.92	✓	**C6**	-	105.65	0.893	
C4	-	>250	0.94		**D6**	-	>250	0.985	
D4	-	>250	0.85	✓	**E6**	+	>250	0.932	

Molecule		T_m Shift Model		
ID	Effect	K_D(μM)	R^2	Toxic
E4	-	>250	0.81	
F4	-	>250	0.81	
G4	-	235.37	0.77	
H4	-	>250	0.96	✓
I4	-	>250	0.87	
J4	-	>250	0.97	✓
K4	-	>250	0.84	
L4	+	219.23	0.69	
M4	-	>250	0.86	
N4	-	>250	0.9	
O4	-	>250	0.9	
P4	-	>250	0.832	
Q4	-	110.36	0.771	
R4	-	249.94	0.926	
S4	-	212.88	0.804	
T4	-	230.87	0.841	
U4	-	224.76	0.952	
V4	-	226.97	0.939	
W4	-	>250	0.956	
Y4	-	142.94	0.921	
Z4	-	>250	0.837	
A5	-	>250	0.758	
B5	-	>250	0.917	✓
C5	-	>250	0.896	
D5	-	121.9	0.802	
F6	-	>250	0.91	
G6	-	>250	0.824	
H6	-	>250	0.817	
I6	-	>250	0.922	
J6	-	116.45	0.855	
K6	-	>250	0.933	
L6	-	101.85	0.729	✓
M6	+	>250	0.732	
N6	-	>250	0.837	
O6	-	>250	0.96	
P6	-	>250	0.917	
Q6	+	>250	0.779	
R6	-	>250	0.843	
S6	-	>250	0.868	
T6	-	>250	0.922	
U6	-	>250	0.781	
V6	+	>250	0.777	
W6	-	>250	0.847	
Y6	-	>250	0.756	
Z6	-	>250	0.868	
A7	-	242.26	0.828	
B7	+	>250	0.815	
C7	-	>250	0.954	
D7	-	132.7	0.911	✓

Supplementary Table 8. Crystallization.

Data-collection and refinement statistics.

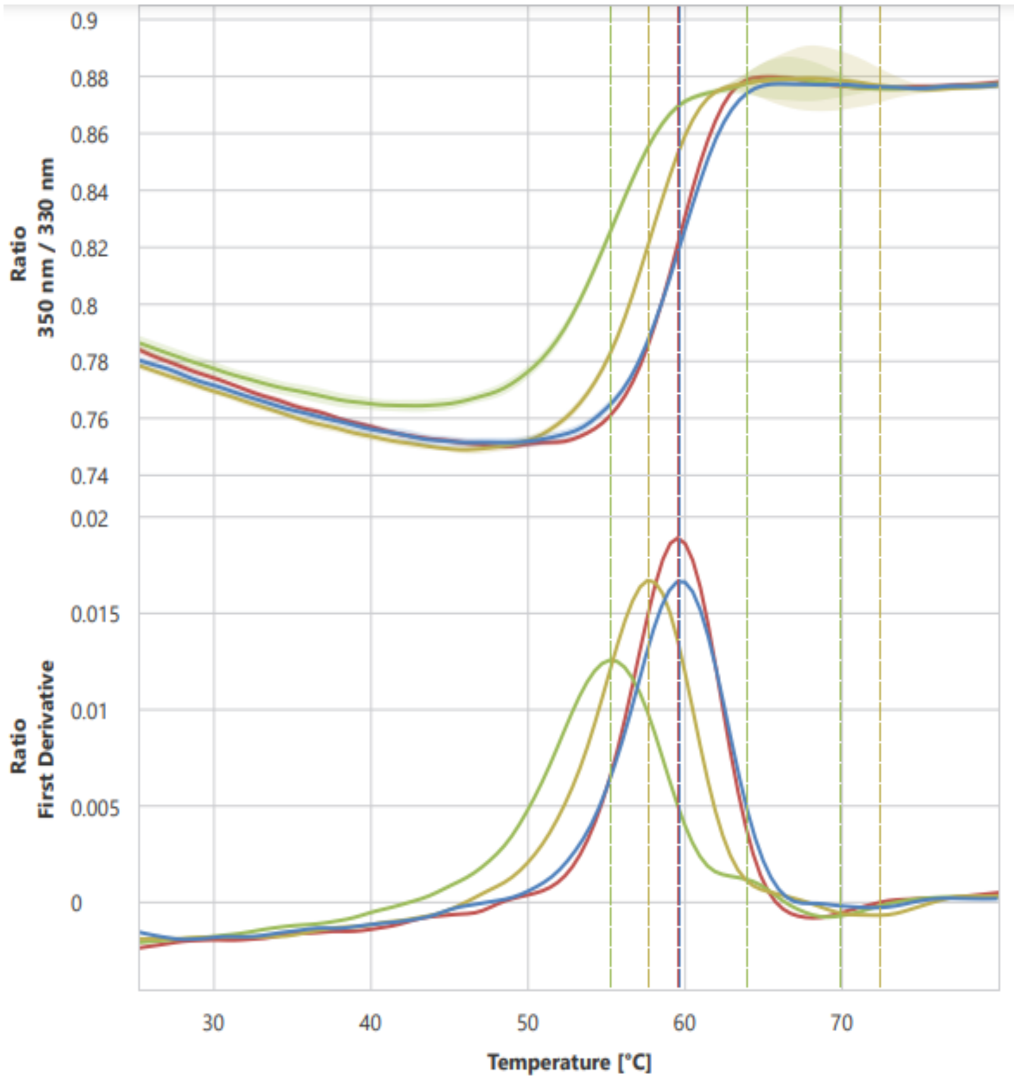
Data collection		**Refinement**	
Beamline	ALBA-BL13	Resolution (Å)	54.82-**1.15**
Wavelength (Å)	0.9793	Reflections	30577
Space group	C 1 2 1	Reflections used for R_{free}	1565
a, *b*, *c* (Å)	113.71, 32.96, 40.11	R_{work} / R_{free}	0.155 / 0.188
α, β, γ (°)	90.00, 105.38, 90.00	No. of non-H atoms	1108
Resolution (Å)*	38.67 - 1.14	Macromolecules	952
	(1.30 - 1.14)	Ligands	3
Total reflections	297645 (14124)	Solvent	153
Unique reflections	30609 (1531)	Protein residues	54
R_{meas}	0.072	DNA base pairs	12
	(1.366)	Average B factors	22.61
$R_{\mathrm{p.i.m}}$	0.023	Macromolecules	19.50
	(0.446)	Ligands	15.72
$I/\sigma(I)$	14.8 (1.8)	Solvent	29.90
$CC_{1/2}$	0.999 (0.611)	RMSD	
Completeness (%):		Bond lengths (Å)	0.013
spherical	59.0 (9.6)	Bond angles (°)	1.35
ellipsoidal#	91.7 (53.3)	Clashscore	0.00
Multiplicity	9.7	Ramachandran %:	
		favored	100
		outliers	0

*Values in parentheses are for the highest resolution shell.

#Anisotropy correction by STARANISO/autoPROC

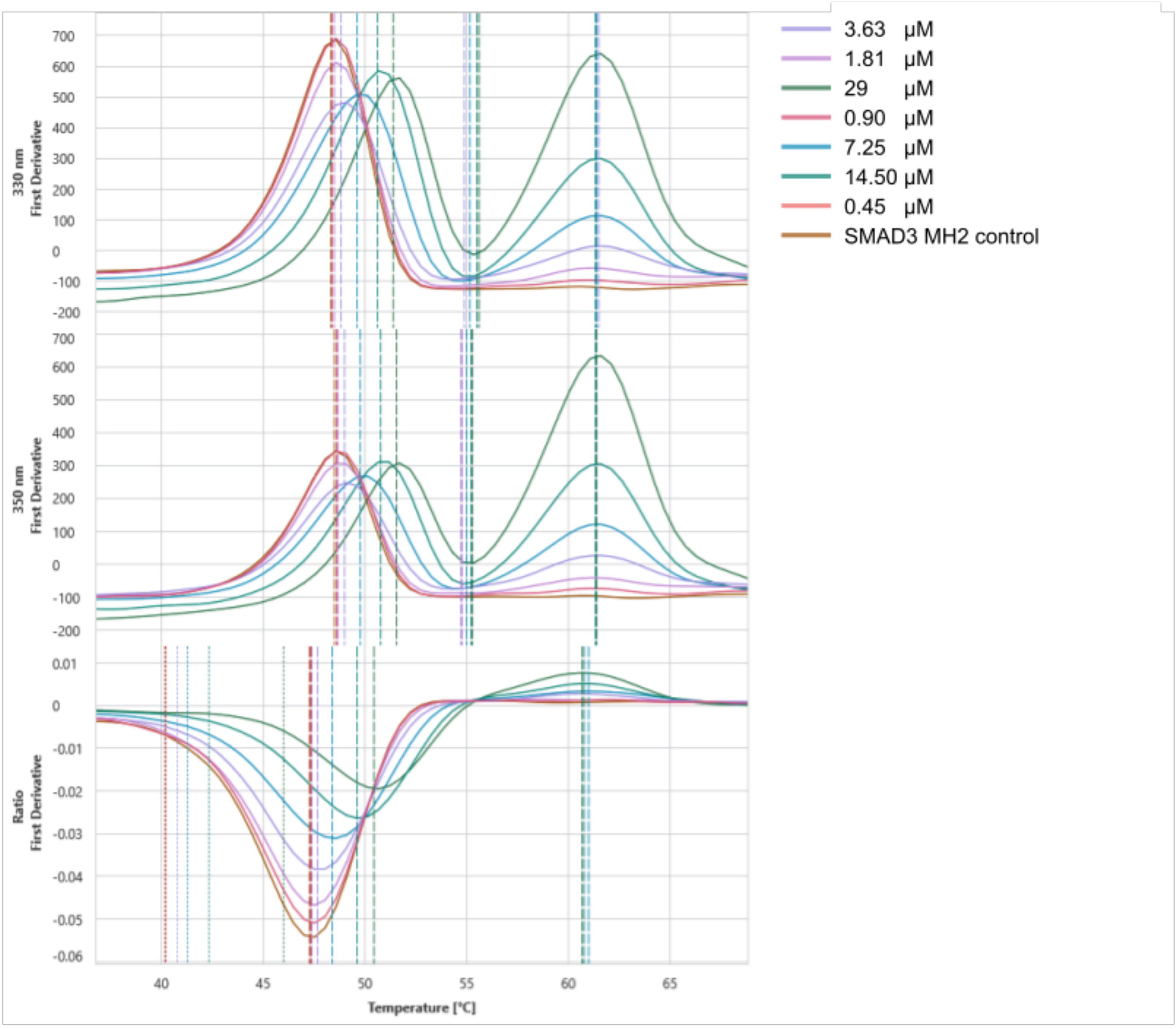
Annex B. Supplementary Figures

Supplementary Figure 1



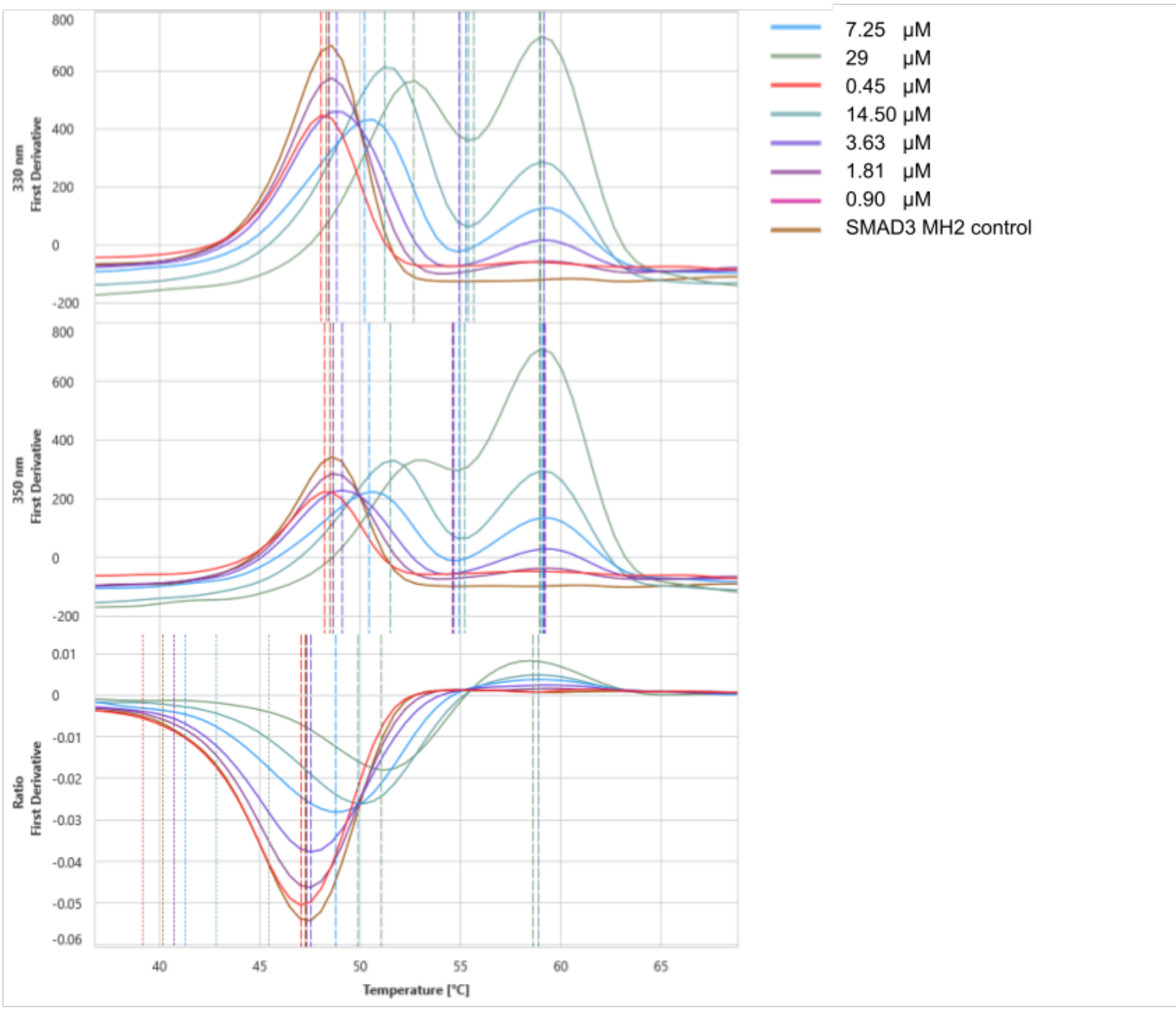
Unfolding profile of SMAD4 MH2 domain variants determined by NanoDSF. Fluorescence ratio 350/330 nm data (TOP) and first derivative (BOTTOM). Data for WT (blue), I500V (red), R496C (yellow) and R361G (green) are shown.

Supplementary Figure 2



NanoDSF unfolding profile of activated SMAD3 MH2 domain mixed with increasing concentrations of SMAD4 MH2 WT. SMAD3 189-425 DVD was used at constant 14.5 µM concentration, SMAD4 MH2 domain concentration is shown in the figure.

Supplementary Figure 3



NanoDSF unfolding profile of activated SMAD3 MH2 domain mixed with increasing concentrations of SMAD4 MH2 R496C. SMAD3 189-425 DVD construct was used at constant 14.5 µM concentration, SMAD4 MH2 domain concentration is shown in the figure.

Annex C. Publications and pre-prints

Publications related to this thesis:

Published, as a co-author:

- HTSDSF Explorer, A Novel Tool to Analyze High-throughput DSF Screenings

  Pau Martin-Malpartida, Emil Hausvik, Jarl Underhaug, **Carles Torner**, Aurora Martinez, Maria J Macias.

  J Mol Biol. 2022 Jun 15;434(11):167372. doi: 10.1016/j.jmb.2021.167372. Epub 2021 Nov 19.

- TPPU_DSF: A web application to calculate thermodynamic parameters using DSF data

  Pau Martin-Malpartida, **Carles Torner**, Aurora Martinez, Maria J. Macias

  J Mol Biol. 2024, In Press.

- AI is a Viable Alternative to High Throughput Screening: a 318-Target Study

  This is a collaborative project coordinated by Atomwise Inc, a company based in the USA and a consortium of more than 300 collaborators from research centers and universities.

  Scientific Reports volume 14, Article number: 7526 (2024)

Publications under review, as a co-author

- TGF-β and RAS jointly unmask primed enhancers to drive metastasis

  Jun Ho Lee, Francisco J. Sánchez-Rivera, Lan He, Harihar Basnet, Fei Chen, Elena Spina, Liangji Li, **Carles Torner**, Jason E. Chan, Dig Vijay Kumar Yarlagadda, Jin Suk Park, Carleigh Sussman, Charles M. Rudin, Scott W. Lowe, Tuomas Tammela, Maria J. Macias, Richard P. Koche, and Joan Massagué

  Under review in Cell

Publications in preparation (two as co-first author)

- Insights into structure-activity relationship of Myhre syndrome variants in SMAD4

  Carles Torner†, Miriam Condeminas†, Radoslaw Pluta, Eric Aragón, Pau Martin-Malpartida, Aurora Martinez and Maria J. Macias (†: Co-first authors)

- Molecular basis for the DNA binding recognition of the Ras-responsive element-binding protein 1 (RREB1)

  Radoslaw Pluta†, Eric Aragón†, **Carles Torner**†, Ingrid Benza, Rebeca A. Mees, Joan Pous, Pau Martin-Malpartida, Jun Ho Lee, Joan Massagué and Maria J. Macias

  (†: Co-first authors)

Publications not included in the PhD thesis

- Unveiling the dimer/monomer propensities of SMAD MH1-DNA complexes.

  Ruiz L, Kaczmarska Z, Gomes T, Aragon E, **Torner C,** Freier R, Baginski B, Martin-Malpartida P, de Martin Garrido N, Marquez JA, Cordeiro TN, Pluta R, Macias MJ.

  Comput Struct Biotechnol J. 2021 Jan 6;19:632-646. doi: 10.1016/j.csbj.2020.12.044. eCollection 2021.

- Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneer-directed TGF-β signaling.

  Aragón E, Wang Q, Zou Y, Morgani SM, Ruiz L, Kaczmarska Z, Su J, **Torner C**, Tian L, Hu J, Shu W, Agrawal S, Gomes T, Márquez JA, Hadjantonakis AK, Macias MJ, Massagué J.

  Genes Dev. 2019 Nov 1;33(21-22):1506-1524. doi: 10.1101/gad.330837.119. Epub 2019 Oct 3.

- TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF-β signaling.

  Guca E, Suñol D, Ruiz L, Konkol A, Cordero J, **Torner C**, Aragon E, Martin-Malpartida P, Riera A, Macias MJ.

  Nucleic Acids Res. 2018 Sep 28;46(17):9220-9235. doi: 10.1093/nar/gky680.

HTSDSF Explorer, A Novel Tool to Analyze High-throughput DSF Screenings

Pau Martin-Malpartida [1*], Emil Hausvik [2], Jarl Underhaug [3], Carles Torner [1], Aurora Martinez [2] and Maria J. Macias [1,4*]

*1 - **Institute for Research in Biomedicine,** The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona 08028, Spain*

*2 - **Department of Biomedicine,** University of Bergen, Jonas Lies vei 91, 5009 Bergen, Norway*

*3 - **Department of Chemistry,** University of Bergen, Allégaten 41, 5007 Bergen, Norway*

*4 - **Institució Catalana de Recerca i Estudis Avançats (ICREA),** Passeig Lluís Companys 23, Barcelona 08010, Spain*

Correspondence to Pau Martin-MalpartidaMaria J. Macias: Institute for Research in Biomedicine, The Barcelona Institute of Science and Technology, Baldiri Reixac, 10, Barcelona 08028, Spain. *pau.martin@irbbarcelona.org (P. Martin-Malpartida), maria.macias@irbbarcelona.org (M.J. Macias), @medofak_uib, @biorecognition (A. Martinez), @maciaslab1 (M.J. Macias)*

https://doi.org/10.1016/j.jmb.2021.167372

Edited by Rita Casadio

Abstract

The identification of new drugs for novel therapeutic targets requires the screening of libraries containing tens of thousands of compounds. While experimental screenings are assisted by high-throughput technologies, in target-based biophysical assays, such as differential scanning fluorimetry (DSF), the analysis steps must be calculated manually, often combining several software packages. To simplify the determination of the melting temperature (T_m) of the target and the change induced by ligand binding (ΔT_m), we developed the HTSDSF explorer, a versatile, all-in-one, user-friendly application suite. Implemented as a server-client application, in the primary screenings, HTSDSF explorer pre-analyzes and displays the T_m and ΔT_m results interactively, thereby allowing the user to study hundreds of conditions and select the primary hits in minutes. This application also allows the determination of preliminary binding constants (K_D) through a series of subsequent dose–response assays on the primary hits, thereby facilitating the ranking of validated hits and the advance of drug discovery efforts.

Introduction

In recent years, the discovery of novel molecules with pharmacological applications has been accelerated thanks to the use of high-throughput screening (HTS) assays that can scan libraries with thousands of molecules each day.[1]

Some of these libraries are organized and distributed to users through actions such as the EU-OPENSCREEN.[2] Supported by HTS platforms throughout Europe, this initiative provides access to a rationally selected compound collection of up to 140,000 commercial and proprietary compounds. Among the strategies used to identify binders for therapeutic targets, differential scanning fluorimetry (DSF) has gained recognition as an affordable and efficient HTS technique to discover innovative candidates in drug discovery projects.[3] The screening is based on the identification of low molecular weight ligands through changes in protein thermal stability upon binding. It is performed using a real-time polymerase chain reaction (RT-PCR) system and fluorescent dyes, such as SYPRO Orange in the case of soluble proteins. The dye binds to hydrophobic patches of the protein that become exposed upon thermal denaturation.[3–6] The same

DSF technique and equipment can be used to acquire a series of dose–response assays (DRAs), thus allowing the determination of preliminary binding constants (K_D) with values often comparable to those obtained by isothermal titration calorimetry[7] and surface plasmon resonance (SPR).[8] These apparent K_D values help categorize the hits prior to performing other orthogonal validation strategies, thus contributing to the hit-to-lead optimization phase.

In most laboratories, the first part of the screening process, which is related to compound and protein handling, is highly automated thanks to the use of 96- or 384-well plates and pipetting robots as well as well-established protocols for protein expression and purification in mg scale (Figure 1 (A)). These automated steps facilitate the reproducibility of the screening and replicates, and the comparison of results between laboratories and users. However, the capacity to screen large libraries of compounds quickly generates huge amounts of data, with the analysis step being one of the main bottlenecks of drug screening. To speed data analysis, several tools have been developed in the last decade. Most tools focus on studying protein stability under different buffer conditions or in low to medium range screening assays.[9–12] However, they do not incorporate information of the molecules nor do they combine the results in a single file displaying the final ranking of best compounds. In addition, hit validation is not included, and this process requires the use of additional general-purpose data analysis software like GraphPad Prism (www.graphpad.com) or OriginLab (originlab.com) to determine the apparent K_D values. To simplify the analysis of large HTS datasets and DRAs in a systematic and user-friendly manner, we developed an open-source package named HTSDSF explorer (Figure 1(B)).

Results

Program description

The HTSDSF explorer is designed as a server-client application that runs locally. The server is coded in python3 and implemented as a custom webserver, which can be downloaded at https://github.com/maciaslab/htsdsf_explorer. The software is compatible with DSF data acquired using different qPCR systems (LightCycler, BioRad and QuantStudio formats), and either 96- or 384-well plates. DSF data are stored as data points containing the fluorescence value and the associated temperature. A ready-to-use version for MS Windows that does not require the installation of python3 is also available.

The client is a web app, coded in JavaScript, which is executed in the browser at the user end. This approach ensures high efficiency and compatibility, as most computing devices and operating systems have a python3 port available and a modern web browser. The server is responsible for loading and managing the DSF data acquired regardless of the qPCR system used, and for converting them into the internal data model that is sent to the client for displaying. The display is user-friendly and highly intuitive. Both the server and client communicate using standard HTTP requests, and data are interchanged using the JavaScript Object Notation (json) format. The server is also responsible for storing persistent data and generating the reports requested by the user, including hit ranking and K_D calculation for hit compounds. A description of the software is included in the accompanying video and in the documentation provided with the package.

Experimental design for HTS binding assays

As an example of a HTS experiment, the screening of 100,000 compounds generates about 300 × 384-well plates and 120,000 experiments to examine, including references. The analysis requires the definition of a threshold for the assay response, which might need to be modified, along with the number of conditions analyzed and the observed melting temperature (T_m) of the target protein without the compound (reference wells or DMSO controls) and with compound, as well as the ΔT_m (T_m with compound − averaged T_m in reference wells. For instance, after analyzing 25,000 compounds and the hits observed, the user might need to increase or decrease the threshold and re-score all the compounds. Also, when manually analyzed, the user has to go through each of the 120,000 experiments, define proper and unique signals, thereby excluding experimental artifacts, and finally, select the list of preliminary hits. In our case, after studying approximately 60,000 compounds provided by EU-OPENSCREEN, we obtained a list of more than 500 promising hits that perturb the target T_m by ±1 °C in the HTS assay. The cutoff value depends on the SD determined in the DMSO control (usually, ±1 °C $\geq$ 3-fold SD). The molecules that destabilize or stabilize the target (the latter also known as pharmacological chaperones in our case study) then need further verification at lower concentrations. This process is typically performed as a dilution series, with the aim to obtain an indicator to rank the hits on the basis of affinity. The results are collected in the form of a K_D score.[13] In our case, this represented 30 additional 384-well plates and 11,000 conditions and the determination of the corresponding 500 K_Ds.

Data analysis

T_m screening

The program starts by displaying a list of files associated with the experimental plates. These files contain the raw experimental data (melting

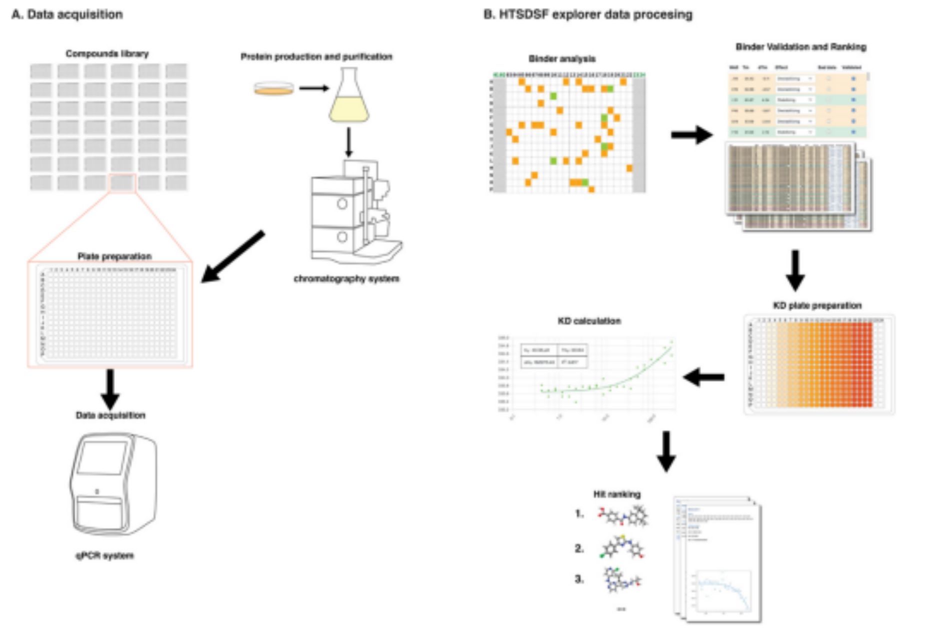
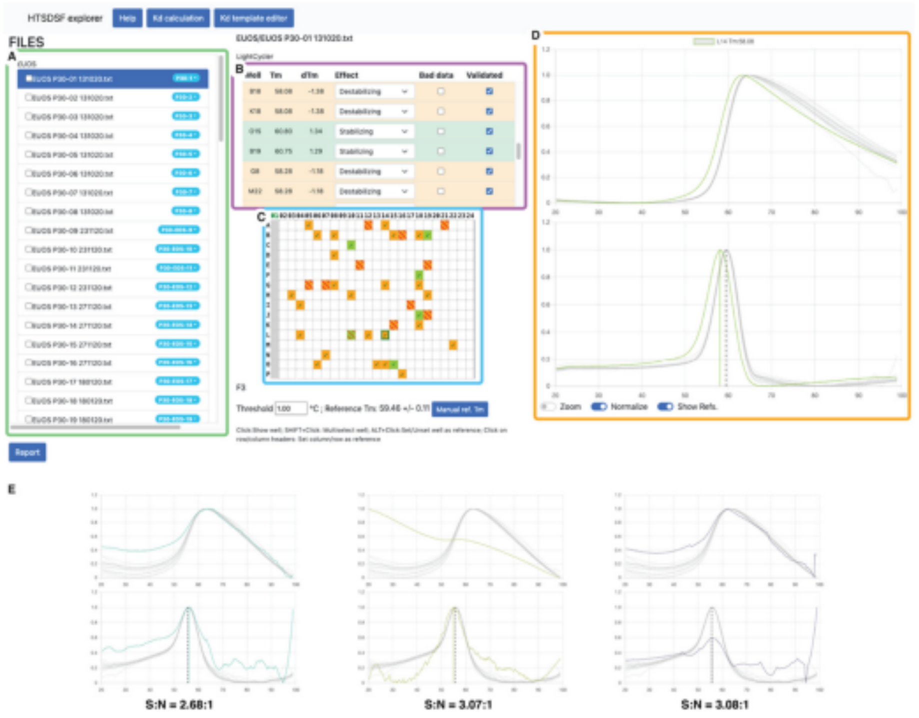


Figure 1. Experimental workflow. (A) Schematic representation of HTS assays. (B) Interface and outputs generated by HTSDSF explorer.

curves of Fluorescence (F) vs. temperature (T) for each well). If the plates belong to a defined library, each file can be correlated to a "plate name" containing information about the compounds dispensed per well. Once a plate is selected from the list (Figure 2(A)), the browser starts showing the results as a table, including the T_m and the ΔT_m (dT_m in Figure 2(B)) for each well, and as an interactive representation of the plate. Both representations simultaneously allow the selection of a given condition. In both, the conditions with an effect on the T_m are highlighted in green (ΔT_m higher than the threshold) or orange (lower) (Figure 2(B, C)). By default, the threshold is set at ±1 °C with respect to the reference target, but the cutoff value can be modified by the user to better fit the temperature changes observed for each specific protein target. Typically, we selected the cutoff at ΔT_m = 5-fold the SD of the T_m of the reference wells. The software supports any number of arbitrarily defined reference wells. In the example, 64 wells were used (shown as two gray columns in Figure 2(B)). When a well or row is selected, the corresponding melting curve and the first derivative are displayed as an additional panel for visual inspection (Figure 2(D)). The determination of the T_m is robust, and works even at signal-to-noise ratios as low as 3:1 (Figure 2 (E)). The user can either validate the T_m-value or flag it as poor or uncertain data. Once this has been done, the next well is automatically loaded and the user repeats the validation procedure until all preselected wells have been evaluated. The validation takes less than five seconds and requires only one mouse-click per well. After validation, the program generates an Excel report ranking all molecules from all selected plates on the basis of T_m changes. The Excel report contains values and the curves for all wells, allowing the user to visualize the curves in the file without the need to go back and forth to the program. For the best conditions, the user can prepare high quality plots from the DSF data.

Hit validation and K_D determination

Once some compounds have been identified as potential hits, it is advisable to prepare DRA plates to estimate apparent K_D values, since the primary screenings are normally performed at high compound excess (e.g., 250 μM). DRA plates are normally designed with the concentrations of the compounds varying along a given row and with an arbitrary number of experimental replicates (Figure 3). The software includes a dose–response plate editor.



EUOS P30-01 131020.txt
EUOS P30-02 131020.txt
EUOS P30-03 131020.txt
EUOS P30-04 131020.txt
EUOS P30-05 131020.txt
EUOS P30-06 131020.txt
EUOS P30-07 131020.txt
EUOS P30-08 131020.txt
EUOS P30-09 231120.txt
EUOS P30-10 231120.txt
EUOS P30-11 231120.txt
EUOS P30-12 231120.txt
EUOS P30-13 271120.txt
EUOS P30-14 271120.txt
EUOS P30-15 271120.txt
EUOS P30-16 271120.txt
EUOS P30-17 180120.txt
EUOS P30-18 180120.txt
EUOS P30-19 180120.txt

EUOS/EUOS P30-01 131020.txt

LightCycler

Well	Tm	dTm	Effect	Bad data	Validated
B18	58.08	−1.38	Destabilizing	□	☑
K18	58.08	−1.38	Destabilizing	□	☑
O15	60.80	1.34	Stabilizing	□	☑
B19	60.75	1.29	Stabilizing	□	☑
G8	58.28	−1.18	Destabilizing	□	☑
M22	58.28	−1.18	Destabilizing	□	☑

F3

Threshold 1.00 °C ; Reference Tm: 59.46 +/- 0.11 Manual ref. Tm

Click: Show well; SHIFT+Click: Multiselect well; ALT+Click: Set/Unset well as reference; Click on row/column headers: Set column/row as reference

Figure 2. Plate and well browser, with the different components highlighted. (A) File browser, (B) Well table, (C) Plate representation, (D) Fluorescence vs. Temperature with calculation of T_m-values. A video showing these features is included. (E) Robust T_m estimation in unfavorable experimental cases with poor signal-to-noise ratios (S:N), S:N is calculated as the mean of the data divided by the standard deviation (SD).

The DR plate-designs stored in the editor can be loaded into the K_D module for K_D calculation. The program will fit the T_m-values for each concentration to the equation described in the methods section.[6,14] In the K_D module, the user can easily disable outliers by clicking on the graph points, and the K_D and ΔH_0 estimations are automatically recalculated after each modification.

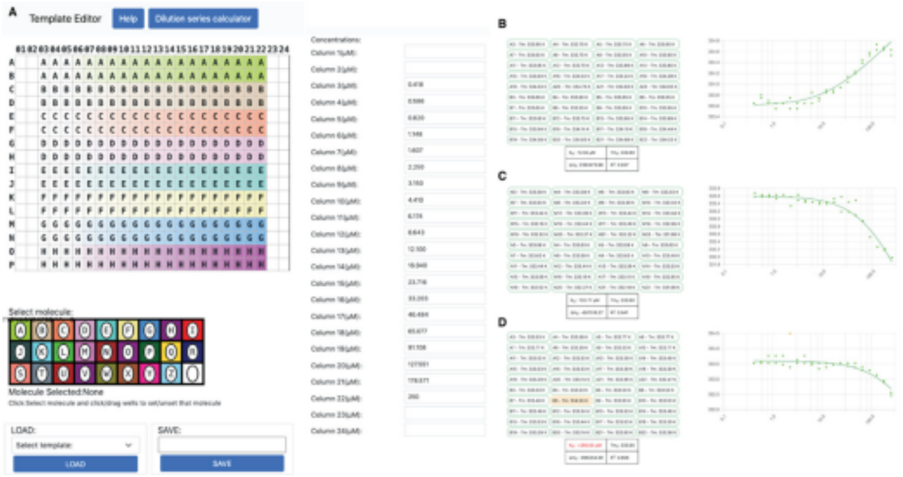
Conclusion

HTSDSF explorer is an all-in-one open-source application suite able to analyze HTS DSF data in a highly intuitive and rapid manner. This software reads input files acquired in the most common qPCR systems, and the data are visualized through a user-friendly interface that allows the user to customize conditions for the analysis and validate the results. HTSDSF explorer has a web interface, but it is run locally, ensuring its reliability and quick access to large amounts of data. The output is a report containing the main features of the experiments (Excel tables and graphs) and it can include either single plate or multiple plate analyses. The same software is also able to design and analyze dose–response assays rapidly and easily and determine apparent K_D constants to consistently categorize hits, allowing to start defining potential pharmacophores. The comparison of chemical properties of hits with other tested molecules belonging to the library is also advantageous in the preparation of the pharmacophores and clustering of compounds. This comparison aids to reduce the number of compounds to be validated and optimized in other expensive and time-consuming orthogonal *in vitro* assays.

Methods

Expression and purification of the protein

Human SMAD4 MH2 domain (272–552) was cloned using an 'In Fusion Cloning strategy'. The



	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A			A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		
B			A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A		
C			B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B		
D			B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B		
E			C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
F			C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C		
G			D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D		
H			D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D		
I			E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E		
J			E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E	E		
K			F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F		
L			F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F		
M			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
N			G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G		
O			H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H		
P			H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H		

Select molecule:

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	( )

Molecule Selected:None

Click Select molecule and click/drag wells to set/unset that molecule

LOAD: Select template: LOAD

SAVE: SAVE

Concentrations:

Column 1(µM):	
Column 2(µM):	
Column 3(µM):	0.418
Column 4(µM):	0.586
Column 5(µM):	0.820
Column 6(µM):	1.148
Column 7(µM):	1.607
Column 8(µM):	2.250
Column 9(µM):	3.150
Column 10(µM):	4.410
Column 11(µM):	6.174
Column 12(µM):	8.643
Column 13(µM):	12.100
Column 14(µM):	16.940
Column 15(µM):	23.716
Column 16(µM):	33.203
Column 17(µM):	46.484
Column 18(µM):	65.077
Column 19(µM):	91.108
Column 20(µM):	127.551
Column 21(µM):	178.571
Column 22(µM):	250
Column 23(µM):	
Column 24(µM):	

Figure 3. (A) Experimental design for an 8-molecule dose–response assay (DRA) with duplicates. (B) Outputs of the K_D module. Stabilizer hit compound. (C) Destabilizer hit compound. (D) Low affinity destabilizer compound, where the maximum concentration is far from saturation.

insert was synthesized by Thermo Fisher Scientific. Codons were optimized for expression in *E. coli* using LB medium at 37 °C. Protein expression was induced with IPTG (0.5 mM) and after induction, the bacteria cultures were incubated O/N at 20 °C. Cultures were centrifuged at 3500*g* for 15 min at 4 °C and the pellet was resuspended in a "lysis buffer" containing 50 mM Tris pH8.0, 400 mM NaCl, 400 mM Imidazole, 0.1 % Tween, and 1 mM TCEP. Protein was purified following standard procedures essentially as described.[15,16] The MH2 domain was further purified by size exclusion chromatography using a preparative grade HiLoad™ 16/60 Superdex75 from GE Healthcare and then concentrated at 7–10 mg/mL, in 20 mM pH 7.5, 100 mM NaCl, 2 mM TECP buffer. We purified ~150 mg of protein for the screening. Protein preparations were verified by Mass Spectrometry and characterized by NMR and SAXS (BMRB: 50737; SASBDB: SASDKG9).[16]

High-throughput screening

The initial HTS step by DSF was performed essentially as described in[6]. Briefly, the experiments were performed with the purified SMAD4 MH2 domain with and without compounds, in a LightCycler 480 Real-Time PCR System (Roche Applied Science), using a total volume of 10 µL in 384-well microplates (Roche Applied Science). Protein was diluted to 50 µM in 20 mM Tris pH 7.5, 100 mM NaCl, and 2 mM TCEP, with 5X SYPRO Orange. Binding results were exported to txt format for analysis with HTSDSF Explorer. Compounds were dissolved in DMSO and then added to the protein and SYPRO Orange solution to a final concentration of 80 µg/mL (corresponding to an averaged compound concentration of 200 µM) and 4% DMSO. Samples were incubated at room temperature for at least 10 min before loading into the PCR-instrument. Controls with 4% DMSO were performed on each plate. Unfolding curves were registered from 20 °C to 95 °C at a scan rate of 2 °C/min.

Accepted data formats

The software has a user-definable data directory to collect the DSF files. HTSDSF Explorer accepts data exported from Bio-Rad (.xlsx), Roche LightCycler (.txt), Applied Biosystems QuantStudio and StepOnePlus (.txt).

In addition, a generic dsf file format (.gdsf) has been defined to allow the use of data acquired in different instruments. This format is a simple text file, with three columns separated by spaces containing, in order, the well, temperature and fluorescence. Details about the file formats can be found at the project web-page.

As the software is open-source, additional formats can be implemented by the user or by us upon request.

Data processing

Melting curves are used to obtain the T_m by calculating the gradient using the *numpy* gradient function and smoothed using a Savitzky-Golay filter.[17] This curve is used to find local maxima (peak picking), which correspond to the protein T_m. This procedure generates curves that are easier to understand than the row data, without altering the T_m.[11] Peak picking is performed using the *scipy* find_peaks function, with prominence = 30% of the vertical curve range. In cases where the compound induces multiple observable transitions, we select the temperature that is closest to the reference T_m value.

Data export and storage

All user-validated T_m data are locally stored in a human-readable format that can be exported to Excel. Reports can either be generated for each plate or for various plates combined as a final report file.

The software allows the user to correlate each well in each plate with a ligand using plate-template information. For this feature, the user needs to fill in a text file with information about the plate ID, the well, the molecule ID and a smiles/InChI string. This information, if available, is added to the report.

K_D calculation

K_D calculation is performed by fitting the data points (ligand concentration and the corresponding T_m) to the equation described in[6].

$T_{m,l} = \frac{\frac{\Delta H_0}{n} T_{m,0}}{\frac{\Delta H_0}{n} - RT_{m,0} \ln\left(\frac{K_d + [L]}{K_d}\right)}$ where $T_{m,l}$ is the melting temperature at a concentration of a given ligand ([L]), K_D is the dissociation constant, $T_{m,0}$ is the melting temperature in the absence of ligand, ΔH_0 is the enthalpy of the unfolding of the protein at $T_{m,0}$, n is the number of binding sites, and R is the gas constant. $T_{m,0}$, ΔH_0 and K_D are obtained after the curve fitting and n is assumed to be 1. Starting values for the fitting are obtained by differential evolution[18] as implemented in *scipy*, and then using these values as initial conditions for a least-squares fitting.[19]

Video link:

http://maciasnmr.net/HTSDSF/HTSDSFvideo.mp4

CRediT authorship contribution statement

Pau Martin-Malpartida: Conceptualization, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Emil Hausvik:** Investigation, Formal analysis. **Jarl Underhaug:** Software. **Carles Torner:** Investigation, Formal analysis. **Aurora Martinez:** Conceptualization, Supervision, Writing – review & editing, Funding acquisition. **Maria J. Macias:** Conceptualization, Supervision, Writing – original draft, Writing – review & editing, Funding acquisition.

Acknowledgements

We acknowledge the support received from the EU-OPENSCREEN team, the BiSS core facility (Department of Biomedicine, University of Bergen (UiB)) and the UiB-node of NOR-OPENSCREEN. We also thank Dr. N. Berrow (IRB Barcelona, Protein Expression Unit) for providing reagents and helping with protein purification.

Funding

The HTS work was supported by EU-OPENSCREEN INSTRUCT-ERIC through funding from the European Union's Horizon 2020 Research and Innovation programme, grant agreement No 823893 (EU-OPENSCREEN-DRIVE). Access to the NMR facility (Grenoble-FR2) was granted by the INSTRUCT-Eric Integrating Biology program (grant 7192, title: Smad4 MH2 domain) and access to Bio-SAXS BM29 was part of the MX-1941 BAG proposal. The work was also financed through the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) (BFU2017-82675-P, M.J.M), IRB Barcelona, AGAUR (SGR-50, M.J.M), the Research Council of Norway (NOR-OPENSCREEN 245922/F50 and FRIMEDBIO 261826, A.M.) and the Western Norway Regional Health Authorities (project 912246 to A.M.). C.T. is a recipient of a PhD fellowship from the Government of Catalonia (Generalitat de Catalunya) (FI) (2021, ref. BDNS 525925). We gratefully acknowledge institutional funding from the CERCA Programme of the Government of Catalonia and from MINECO through the Centres of Excellence Severo Ochoa Award. M.J.M is an ICREA Programme Investigator.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received 1 October 2021;
Accepted 15 November 2021;
Available online 19 November 2021

***Keywords*:**
high-throughput compound screening;
thermal shift assay;
differential scanning fluorimetry;
automatic T_m;
automatic K_D analysis

References

1. Macarron, R., Banks, M.N., Bojanic, D., Burns, D.J., Cirovic, D.A., Garyantes, T., et al., (2011). Impact of high-throughput screening in biomedical research. *Nature Rev. Drug Discov.* **10**, 188–195.
2. Brennecke, P., Rasina, D., Aubi, O., Herzog, K., Landskron, J., Cautain, B., et al., (2019). EUOPENSCREEN: A novel collaborative approach to facilitate chemical biology. *SLAS Discov.* **24**, 398–413.
3. Niesen, F.H., Berglund, H., Vedadi, M., (2007). The use of differential scanning fluorimetry to detect ligand interactions that promote protein stability. *Nature Protoc.* **2**, 2212–2221.
4. Boyd, R.E., Lee, G., Rybczynski, P., Benjamin, E.R., Khanna, R., Wustman, B.A., et al., (2013). Pharmacological chaperones as therapeutics for lysosomal storage diseases. *J. Med. Chem.* **56**, 2705–2725.
5. Pantoliano, M.W., Petrella, E.C., Kwasnoski, J.D., Lobanov, V.S., Myslik, J., Graf, E., et al., (2001). Highdensity miniaturized thermal shift assays as a general strategy for drug discovery. *J. Biomol. Screen.* **6**, 429–440.
6. Støve, S.I., Flydal, M.I., Hausvik, E., Underhaug, J., Martinez, A., (2020). Differential scanning fluorimetry in the screening and validation of pharmacological chaperones for soluble and membrane proteins, Chapter 15. Academic Press, Walthum.
7. Gao, K., Oerlemans, R., Groves, M.R., (2020). Theory and aapplications of differential scanning fluorimetry in earlystage drug discovery. *Biophys. Rev.* **12**, 85–104.
8. Martin, I., Underhaug, J., Celaya, G., Moro, F., Teigen, K., Martinez, A., Muga, A., (2013). Screening and evaluation of small organic molecules as ClpB inhibitors and potential antimicrobials. *J. Med. Chem.* **56** (18), 7177–7189.
9. Lee, P.H., Huang, X.X., Teh, B.T., Ng, L.M., (2019). TSACRAFT: A Free Software for Automatic and Robust Thermal Shift Assay Data Analysis. *SLAS Discov.* **24**, 606–612.
10. Rosa, N., Ristic, M., Seabrook, S.A., Lovell, D., Lucent, D., Newman, J., (2015). Meltdown: A tool to help in the interpretation of thermal melt curves acquired by differential scanning fluorimetry. *J. Biomol. Screen.* **20**, 898–905.
11. Sun, C., Li, Y., Yates, E.A., Fernig, D.G., (2020). SimpleDSFviewer: A tool to analyze and view differential scanning fluorimetry data for characterizing protein thermal stability and interactions. *Protein Sci.* **29**, 19–27.
12. Wang, C.K., Weeratunga, S.K., Pacheco, C.M., Hofmann, A., (2012). DMAN: A Java tool for analysis of multiwell differential scanning fluorimetry experiments. *Bioinformatics* **28**, 439–440.
13. Bai, N., Roder, H., Dickson, A., Karanicolas, J., (2019). Isothermal analysis of thermofluor data can readily provide quantitative binding affinities. *Sci. Rep.* **9**, 2650.
14. Owen, T., (1994). The origin of inner planet atmospheres. *Philos. Trans. Phys. Sci. Eng.* **349**, 209–211. discussion 12.
15. Aragon, E., Wang, Q., Zou, Y., Morgani, S.M., Ruiz, L., Kaczmarska, Z., et al., (2019). Structural basis for distinct roles of SMAD2 and SMAD3 in FOXH1 pioneerdirected TGFbeta signaling. *Genes Dev.* **33**, 1506–1524.
16. Gomes, T., Martin, P., Malpartida, Ruiz, L., Aragón, E., Cordeiro, T.N., Macias, M.J., (2021). Conformational landscape of multidomain SMAD proteins. *Comput. Struct. Biotechnol. J.* **14** (19), 5210–5224.
17. Savitzky, A., Golay, M.J.E., (1964). Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639.
18. Storn, R., Price, K., (1997). Differential evolution – A simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.* **11**, 341–359.
19. Branch, M.A., Coleman, T.F., Li, Y., (1999). A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. *SIAM J. Sci. Comput.* **21**, 1–23.

TPPU_DSF: A Web Application to Calculate Thermodynamic Parameters Using DSF Data

Pau Martin-Malpartida [1,*], **Carles Torner** [1], **Aurora Martinez** [2] **and Maria J. Macias** [1,3,*]

*1 - **Institute for Research in Biomedicine,** The Barcelona Institute of Science and Technology (BIST), Baldiri Reixac, 10, Barcelona 08028, Spain*

*2 - **Department of Biomedicine,** and the Kristian Gerhard Jebsen Center for Translational Research in Parkinson's Disease, University of Bergen, 5020 Bergen, Norway*

*3 - **Institució Catalana de Recerca i Estudis Avançats (ICREA),** Passeig Lluís Companys 23, Barcelona 08010, Spain*

Correspondence to Pau Martin-Malpartida Maria J. Macias: *maria.macias@irbbarcelona.org (M.J. Macias)*

https://doi.org/10.1016/j.jmb.2024.168519

Edited by Rita Casadio

Abstract

Here we present TPPU_DSF (https://maciasnmr.net/tppu_dsf/). This is a free and open-source web application that opens, converts, fits, and calculates the thermodynamic parameters of protein unfolding from standard differential scanning fluorimetry (DSF) data in an automated manner. The software has several applications. In the context of screening compound libraries for protein binders, obtaining thermodynamic parameters provides a more robust approach to detecting hits than the changes in the melting temperature (T_m) alone, thereby helping to increase the number of positive hits in screening campaigns. Moreover, changes in ΔG_u° indicate protein response to binding at lower compound concentrations than those in the T_m, thereby reducing the costs associated with the amounts of protein and compounds required for the assays. Also, by adding thermodynamic information to the T_m comparison, the software can contribute to the optimization of protein constructs and buffer conditions, a common practice before structural and functional projects.

Introduction

Differential Scanning Fluorimetry (DSF) is a fast and affordable technique to determine protein melting temperature (T_m). Changes in T_m are often used to measure the effects of variations in pH and buffer composition on protein stability. These changes have also been used to detect complex formation when other biomolecules are added, or to indicate small molecule binding when searching for hits with potential applications in drug discovery. Given that DSF assays can be performed in high-throughput formats such as multi-well plates, the technique has gained wide acceptance as a method for the easy and rapid screening of large libraries of compounds. This is especially true when computational tools are available to facilitate the analysis of the results. We had previously developed a computational tool to facilitate the determination of T_m changes during the high throughput screening (HTS) of large libraries of compounds (HTSDSF explorer).[1] Given the large number of compounds analyzed in HTS projects, a key decision in the screening protocol is to establish the ΔT_m (threshold) at which a hit is distinguished from noise. In addition, if several related compounds produce a similar ΔT_m near the threshold, it is difficult to decide which hits to select or discard. A compromise must be made to avoid over- or under-selecting hits, as the selection of many adds complexity and cost to subsequent validation by dose–response assays. We consid-

0022-2836/© 2024 Elsevier Ltd. All rights reserved.

Please cite tis article as: P. Martin-Malpartida, C. Torner, A. Martinez, et al., TPPU_DSF: A Web Application to Calculate Thermodynamic Parameters Using DSF DataJournal of Molecular Biology, https://doi.org/10.1016/j.jmb.2024.168519

ered that the addition of more information, beyond that provided by the $\Delta T_{m,}$ to the hit selection process, would help identify bona fide hits close to the threshold without eliminating potential candidates or including too many non-binders. In fact, the methodology to obtain thermodynamic data from DSF has already been described for proteins undergoing unfolding through a two-state transition.[2] However, in that study, a manual approach was used, thereby limiting its efficient application for HTS projects. Here, we set out to implement TPPU_DSF, a software to determine the thermodynamic properties of unfolding as well as changes in T_m values, to provide a global information on the types of interaction, binding mechanisms and conformational changes that are associated with ligand recognition in an automated manner.

TPPU_DSF is an open-source and easy-to-use web application designed to help generalize the determination of ΔG_u^o, ΔH_u^o, and ΔS_u^o values from DSF experiments. The software is compatible with DSF data obtained from commonly used systems and it is very fast. The analysis of a full plate, starting from data loading to the export of the final results, takes only a few seconds per plate on a conventional personal computer. We are confident that the user-friendly interface of TPPU_DSF will enable many researchers to routinely incorporate thermodynamic information into their experiments. It will also add a second layer of robustness to the compound library screening process using DSF and streamline the optimization of conditions for structural biology studies.

Results

Program description

TPPU_DSF is a web application that automatically opens, converts, fits, and calculates the **T**hermodynamic **P**arameters of **P**rotein **U**nfolding from standard **DSF** data. It is programmed as a fully client-side web application, meaning that all the processing and calculations are executed on the user's computer, even if the application is accessed through a web browser. Therefore, the data are not transferred through the internet, thus ensuring confidentiality. TPPU_DSF can be accessed from https://maciasnmr.net/tppu_dsf/, and its source code is available at https://github.com/maciaslab/tppu_dsf.

A snapshot of the TPPU_DSF user interface with the different components is shown in Figure 1. The interface is based on a multi-step wizard, where different options are displayed sequentially as the user completes the requests. To load data, files can either be selected from the user's file system, or dragged and dropped into the application, and the software automatically detects the file format. There is also a button to load the example data, which can be downloaded as a reference file with a format that works correctly on the server.

The software is compatible with data from the Roche Lightcycler, ThermoFisher StepOne, ThermoFisher Quantstudio, BioRad, and NanoTemper Prometheus systems. Once the data file has been read, the program calculates T_m values (using both the Boltzmann and the derivative method), as well as $\Delta G_u{}^o$, $\Delta H_u{}^o$, and $\Delta S_u{}^o$, for each well (or capillary in the case of the NanoTemper Prometheus). These values are collected as a table. Additionally, the user can manually explore the wells and visualize the denaturation curve and the corresponding thermodynamic parameters. The program also allows the user to simultaneously select several wells and calculate the average and the standard deviation for each thermodynamic parameter, thereby facilitating the analysis of experimental replicates. The user can also export the raw data and the plots for each well in PDF format (with publication-quality displays).

Data analysis

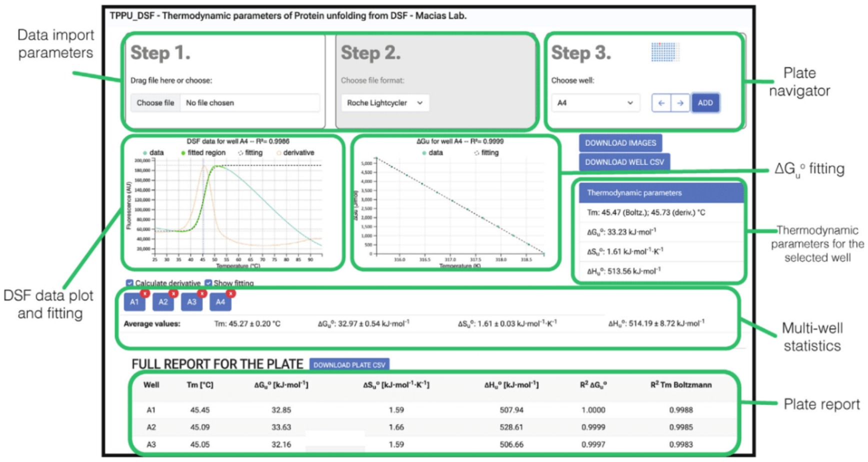
Thermodynamic parameters of unfolding

The calculation of T_m from the thermal denaturation curve is obtained using the two most widely used approaches, namely the Boltzmann and the derivative methods. The first way is based on fitting the DSF thermal unfolding curve to the Boltzmann equation (1):

$$F = F_{min} + \frac{F_{max} - F_{min}}{1 + e^{(T_m - T)/s}} \qquad (1)$$

in which F is the Fluorescence at temperature T, $Fmax$ and $Fmin$ are the maximum and minimum values for the fluorescence, T_m is the melting temperature, and s is the slope of the linear region of the sigmoidal curve. As restrictions to the fitting function, the determination of the minimum fluorescence value should not use the first 10 % of the data points (in which there can be data that do not follow a sigmoidal function because of dye aggregation or other phenomena), and the maximum fluorescence value should occur at higher temperatures than the minimum. The second way to calculate T_m is the derivative method. We show users both methods, as most DSF software packages do, so that they can choose according to their preference since both provide slightly different (but highly comparable) results.

To calculate the thermodynamic parameters of unfolding ($\Delta G_u{}^o$, $\Delta H_u{}^o$, and $\Delta S_u{}^o$), we followed the protocol described by Wright *et al.*[2] First, we estimate the fraction of folded and unfolded protein at each temperature, assuming that the protein is fully folded when $F = Fmin$ and fully unfolded when $F = Fmax$. Using a simple linear interpolation, the fraction of unfolded protein (P_u) at each temperature is given by Eq. (2)



TPPU_DSF - Thermodynamic parameters of Protein unfolding from DSF - Macias Lab.

Step 1. Drag file here or choose: Choose file | No file chosen

Step 2. Choose file format: Roche Lightcycler

Step 3. Choose well: A4 ← → ADD

DOWNLOAD IMAGES

DOWNLOAD WELL CSV

Thermodynamic parameters

Tm: 45.47 (Boltz.); 45.73 (deriv.) °C

$\Delta G_u°$: 33.23 kJ·mol^{-1}

$\Delta S_u°$: 1.61 kJ·mol^{-1}·K^{-1}

$\Delta H_u°$: 513.56 kJ·mol^{-1}

☑ Calculate derivative ☑ Show fitting

A1 A2 A3 A4

Average values: Tm: 45.27 ± 0.20 °C $\Delta G_u°$: 32.97 ± 0.54 kJ·mol^{-1} $\Delta S_u°$: 1.61 ± 0.03 kJ·mol^{-1}·K^{-1} $\Delta H_u°$: 514.19 ± 8.72 kJ·mol^{-1}

FULL REPORT FOR THE PLATE DOWNLOAD PLATE CSV

Well	Tm [°C]	$\Delta G_u°$ [kJ·mol^{-1}]	$\Delta S_u°$ [kJ·mol^{-1}·K^{-1}]	$\Delta H_u°$ [kJ·mol^{-1}]	R^2 $\Delta G_u°$	R^2 Tm Boltzmann
A1	45.45	32.85	1.59	507.94	1.0000	0.9988
A2	45.09	33.63	1.66	528.61	0.9999	0.9985
A3	45.05	32.16	1.59	506.66	0.9997	0.9983

Figure 1. A snapshot of TPPU_DSF showing the different components.

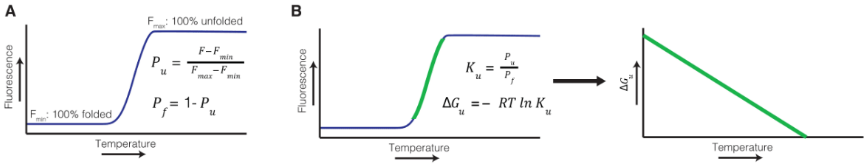


Figure 2. Schematic depiction of the derivation of the fraction of folded protein (A), and ΔG_u and ΔG_u^o (B) from a DSF curve.

$$P_u = \frac{F - F_{min}}{F_{max} - F_{min}} \tag{2}$$

with the fraction of folded protein (P_f) being defined as 1-P_u (Figure 2A). From this, we can obtain the equilibrium constant of unfolding with Eq. (3)

$$K_u = \frac{P_u}{P_f} \tag{3}$$

As ΔG_u is defined as $\Delta_u G = -RTlnK_u$, we can obtain ΔG_u for the linear part of the unfolding curve from K_u. Given that ΔG_u is inversely proportional to the temperature in this region, the linear regression can be used to extrapolate $\Delta G_u{}^o$, the standard Gibbs free energy change of unfolding (ΔG_u at standard ambient temperature and pressure (SATP), T = 298 K and P = 100 kPa) (Figure 2B).

Once $\Delta G_u{}^o$ has been obtained, $\Delta H_u{}^o$, and $\Delta S_u{}^o$ can also be deduced using equations (4) and (5):

$$\Delta S_u{}^o = \frac{\Delta G_u{}^o}{(T_m - T)} \tag{4}$$

$$\Delta H_u{}^o = T_m \cdot \Delta S_u{}^o \tag{5}$$

Output of the analysis

The program can return different output files. All the data used to generate the plots for the DSF data, and the $\Delta G_u{}^o$ fitting can be obtained as a standard CSV file (Microsoft Excel compatible). In addition, the figures can be downloaded as publication-ready and fully editable vectorial PDF files (an example is shown in Supplementary Figure 1). The software is user friendly, and the derivation of thermodynamic parameters of unfolding from DSF data can be easily performed. Potential applications include individual experiments, such as those aimed at optimizing protein constructs or buffer conditions for crystallization experiments, or for functional assays in vitro or as part of HTS campaigns in drug discovery, without adding computational burden to data analysis.

Specific application to HTS

Over the past several years, we have been working to uncover new drug-binding hotspots and compounds as new therapeutic opportunities to treat diseases associated with dysfunction of the TGFβ/SMAD signaling pathway.[3–5] We have focused on analyzing the SMAD4 protein as variants or deletions in the SMAD4 gene have been implicated in pancreatic and colorectal cancer, juvenile polyposis syndrome, and hereditary hemorrhagic telangiectasia.[6,7] Specific mutations are identified in Myhre syndrome, a rare autosomal dominant disorder characterized by skeletal abnormalities, distinctive facial features, intellectual disability, and heart defect. There are currently no effective treatments for this syndrome.[8,9]

To identify SMAD4 binders, we started screening a large library of 100,037 compounds provided by EU-OPENSCREEN.[10] Compact protein domains, and in particular the SMAD4 MH2 domain, belong to the two-state folding/unfolding class,[11,12] and are therefore suitable for using DSF (see methods for details) to monitor the binding of small molecules. To facilitate the identification of binders for such a large library of compounds, we developed the HTSDSF Explorer[1], an application suite that interactively displays T_m and ΔT_m results to explore hundreds of conditions (96/384 well plates) and identify primary hits. Compounds that produced a change in T_m greater than one to five times the standard deviation were considered hits in our study.

As it is possible to extract additional thermodynamic properties, such as Gibbs energy, alongside the evaluation of T_m changes, we reanalyzed the entire primary screening and also the Dose-Response Assays plates of the selected hits with the TPPU_DSF web application. The aim was to determine whether changes in other thermodynamic parameters, such as $\Delta G_u{}^o$, could help identify additional hit candidates that were discarded because they did not fulfill the selection ΔT_m criteria.

In the analysis, we observed that the inflection point in the $\Delta G_u{}^o$ curve was detected at a lower compound concentration compared to the T_m plot. This feature reflects that the sensitivity of $\Delta G_u{}^o$ is enhanced with respect to the T_m value, thus requiring lower compound concentrations for detecting binding. Furthermore, we found that even small changes in T_m upon binding can result in significant variations in $\Delta G_u{}^o$, thus demonstrating the potential of using this thermodynamic parameter as an indicator of binding. Also, in the primary screen, we observed cases where the protein T_m was almost unaffected by the addition of the compound, which instead induced a substantial change in $\Delta G_u{}^o$ (greater than 5 times the standard deviation of the protein's $\Delta G_u{}^o$ in the absence of the compound). We also observed that the dose–response curve for $\Delta G_u{}^o$ was typically sigmoidal, and when plotted against concentration, the exponential decay observed in the T_m curve was less pronounced than for $\Delta G_u{}^o$, as shown in Figure 3. Remarkably, we did not observe a change in $\Delta G_u{}^o$ in cases for which there was no correlation between T_m and compound concentration. This finding indicates that there were no false positive results introduced after the analysis of $\Delta G_u{}^o$ values.

Overall, these results indicate that the pronounced effect of protein ΔG_u^o values on compound binding can provide an internal validation method of the screen and help increase the number of hits during HTS campaigns by identifying compounds that may have been missed when considering only T_m values. The software provides a second advantage in cases where there is a limiting factor in the amount of compound used for the assay because the monitoring of changes in ΔG_u^o requires lower concentrations of ligand than those required to show changes in T_m.

Conclusion

The optimization of experimental conditions for structural biology projects and the identification of new compounds with pharmacological applications are time-consuming and costly processes. To save resources and speed up the analysis step of DSF-based HTS campaigns, we have developed the TPPU_DSF web application, which helps calculate the thermodynamic parameters of protein unfolding based on data obtained from DSF assays.

Methods

Expression and purification of the protein

SMAD4 MH2 domain (Uniprot Q13485, aa 272–552). The domain was expressed in *E. coli* in LB medium at 37 °C, induced with IPTG, and followed by O/N incubation. After cell centrifugation and lysis, the protein was found in the soluble supernatant and purified by nickel-affinity chromatography (HiTrap Chelating HP 5 mL column, GE Healthcare Life Science) using an NGC Quest 10 Plus Chromatography System (BIO-RAD) and eluted with an imidazole gradient. Tag removal was monitored by SDS-PAGE, and cleaved proteins were further purified through size-exclusion chromatography using HiLoadTM Superdex 75 16/60 prep-grade columns (GE Healthcare) equilibrated in buffer 1 (20 mM Tris pH 7.5, 100 mM NaCl, and 2 mM TCEP). Protein purity was confirmed by SDS-PAGE and protein integrity by Mass Spectrometry prior to the binding assays, essentially as described in[13]. NMR, as well as DSF, revealed that the MH2 domain of SMAD4 belongs to the two-state unfolding class.[14]

DSF assays

The initial HTS was performed with the purified SMAD4 MH2 domain with and without compounds. DSF experiments were carried out

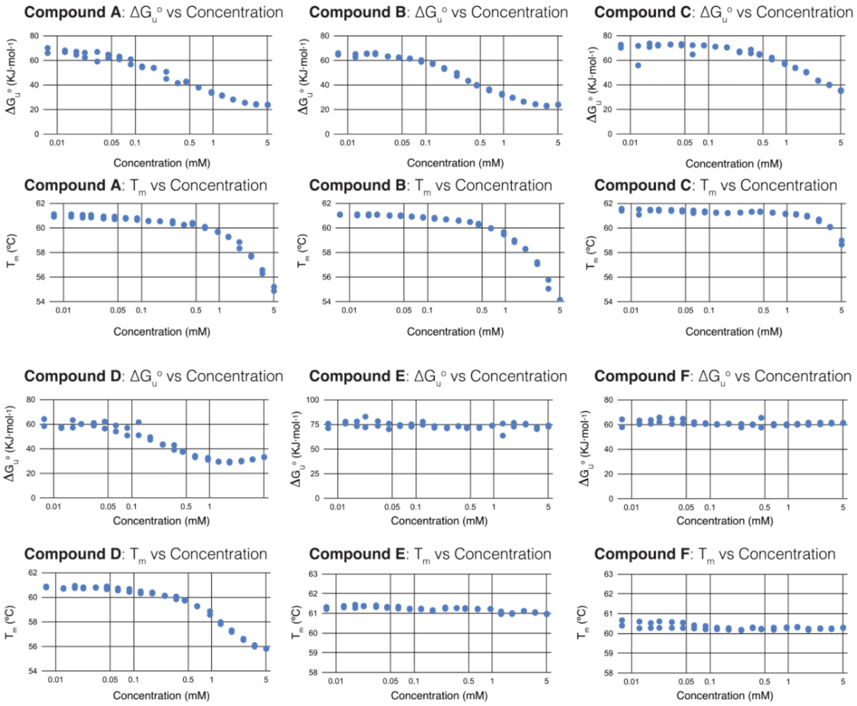


Figure 3. Plots for T_m and $\Delta G_u{}^o$ vs concentration for 4 distinct SMAD4MH2 binders (compounds A-D) and two non-binders (compounds E and F) in a dose–response assay. Data were acquired in duplicates.

using a LightCycler 480 Real-Time PCR System (Roche Applied Science), with a total reaction volume of 10 µL in 384-well microplates (Roche Applied Science). Protein was diluted to 50 µM in buffer 1. SYPRO Orange was used as a dye for the DSF assay, as described in[1].

Software and used libraries

The software was developed using JavaScript. It is compatible with all major browsers, including Mozilla Firefox, Google Chrome, Apple Safari, and Microsoft Edge. We used the *Bootstrap* toolkit to get a basic CSS scaffold [https://getbootstrap.com/]. *fminsearch* was used for the non-linear regression (Boltzmann fit) [https://github.com/jonasalmeida/fminsearch]. ml-savitzky-golay [https://github.com/mljs/savitzky-golay] was used to smooth the DSF curve and calculate its derivative. The d3.js library [https://d3js.org/] was used to plot the curves. PDFKit [https://pdfkit.org/], SVG-to-PDFKit [https://github.com/alafr/SVG-to-PDFKit] and blob-stream [https://github.com/devongovett/blob-stream] were used to generate the PDF file.

Funding

Libraries of compounds were provided by EU-OPENSCREEN INSTRUCT-ERIC, funded by the European Union's Horizon 2020 Research and Innovation programme (grant agreement No 823893) (EU-OPENSCREEN-DRIVE and MedChem). The research was funded by the following projects: PID2021-122909NB-I00 from the Spanish Ministerio de Ciencia e Innovación MCIN/ AEI /https://doi.org/10.13039/501100011033/ and FEDER -A Way of making Europe; PDC_2021-121162-I00 from MCIN and by the European Union Next Generation EU/PRTR; SGR 866-2021 and Producte (2023 PROD_00095) from the AGAUR (M.J.M). This work was also supported by the Stiftelsen K.G. Jebsen and The Neuro-SysMed Center (RCN, Project No. 288164, A.M.). C.T. was awarded a doctoral fellowship from AGAUR. We also acknowledge the institutional support of the Department of Biomedicine, University of Bergen, the IRB Barcelona, the MCIN through the Centers of Excellence Severo Ochoa Award, the BBVA Foundation, and the CERCA program of the Government of Catalonia. M.J.M. is an ICREA Program Investigator.

CRediT authorship contribution statement

Pau Martin-Malpartida: Writing – review & editing, Writing – original draft, Software, Formal analysis, Conceptualization. **Carles Torner:** Investigation, Formal analysis. **Aurora Martinez:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Maria J. Macias:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We greatly appreciate the support of the EU-OPENSCREEN program and its core team for providing compounds and help. Thanks also go to the BiSS core facility (Department of Biomedicine, University of Bergen, Norway), and the UiB-node of NOR-OPENSCREEN (Bergen, Norway). We thank the authors of the software libraries used in the development of TPPU_DSF and the anonymous reviewers for their contributions to the peer review of this work.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2024.168519.

Received 24 January 2024;
Accepted 1 March 2024;
Available online xxxx

Keywords:
Differential scanning fluorimetry;
thermodynamic parameters;
high-throughput compound screening;
thermal shift assay;
hit selection, hit rate

References

1. Martin-Malpartida, P., Hausvik, E., Underhaug, J., Torner, C., Martinez, A., Macias, M.J., (2022). HTSDSF explorer, a novel tool to analyze high-throughput DSF screenings. *J. Mol. Biol.* **434**, 167372
2. Wright, T.A., Stewart, J.M., Page, R.C., Konkolewicz, D., (2017). Extraction of thermodynamic parameters of protein unfolding using parallelized differential scanning fluorimetry. *J. Phys. Chem. Letter* **8**, 553–558.
3. Clackson, T., Wells, J.A., (1995). A hot spot of binding energy in a hormone-receptor interface. *Science* **267**, 383–386.
4. Macias, M.J., Martin-Malpartida, P., Massague, J., (2015). Structural determinants of Smad function in TGF-beta signaling. *Trends Biochem. Sci* **40**, 296–308.
5. Massagué, J., Sheppard, D., (2023). TGF-beta signaling in health and disease. *Cell* **186**, 4007–4037.
6. Wan, R., Feng, J., Tang, L., (2021). Consequences of mutations and abnormal expression of SMAD4 in tumors and T cells. *Onco. Targets Ther.* **14**, 2531–2540.
7. Fang, T., Liang, T., Wang, Y., Wu, H., Liu, S., Xie, L., et al., (2021). Prognostic role and clinicopathological features of SMAD4 gene mutation in colorectal cancer: a systematic review and meta-analysis. *BMC Gastroenterol.* **21**, 297.
8. Alankarage, D., Enriquez, A., Steiner, R.D., Raggio, C., Higgins, M., Milnes, D., et al., (2022). Myhre syndrome is caused by dominant-negative dysregulation of SMAD4 and other co-factors. *Differentiation* **128**, 1–12.
9. Starr, L.J., Grange, D.K., Delaney, J.W., Yetman, A.T., Hammel, J.M., Sanmann, J.N., et al., (2015). Myhre syndrome: clinical features and restrictive cardiopulmonary complications. *Am. J. Med. Genet. A* **167A**, 2893–2901.
10. Brennecke, P., Rasina, D., Aubi, O., Herzog, K., Landskron, J., Cautain, B., et al., (2019). EU-OPENSCREEN: a novel collaborative approach to facilitate chemical biology. *SLAS Discov.* **24**, 398–413.
11. Fersht, A.R., Daggett, V., (2002). Protein folding and unfolding at atomic resolution. *Cell* **108**, 573–582.
12. Zwanzig, R., (1997). Two-state models of protein folding kinetics. *PNAS* **94**, 148–150.
13. Martin-Malpartida, P., Batet, M., Kaczmarska, Z., Freier, R., Gomes, T., Aragon, E., et al., (2017). Structural basis for genome wide recognition of 5-bp GC motifs by SMAD transcription factors. *Nature Commun.* **8**, 2070.
14. Gomes, T., Martin-Malpartida, P., Ruiz, L., Aragon, E., Cordeiro, T.N., Macias, M.J., (2021). Conformational landscape of multidomain SMAD proteins. *Comput. Struct. Biotechnol. J.* **19**, 5210–5224.

scientific reports

OPEN

AI is a viable alternative to high throughput screening: a 318-target study

The Atomwise AIMS Program[1]*

High throughput screening (HTS) is routinely used to identify bioactive small molecules. This requires physical compounds, which limits coverage of accessible chemical space. Computational approaches combined with vast on-demand chemical libraries can access far greater chemical space, provided that the predictive accuracy is sufficient to identify useful molecules. Through the largest and most diverse virtual HTS campaign reported to date, comprising 318 individual projects, we demonstrate that our AtomNet® convolutional neural network successfully finds novel hits across every major therapeutic area and protein class. We address historical limitations of computational screening by demonstrating success for target proteins without known binders, high-quality X-ray crystal structures, or manual cherry-picking of compounds. We show that the molecules selected by the AtomNet® model are novel drug-like scaffolds rather than minor modifications to known bioactive compounds. Our empirical results suggest that computational methods can substantially replace HTS as the first step of small-molecule drug discovery.

Despite present interest in AI/ML and thirty years of case studies[1–4], computational screening techniques have achieved limited adoption within the pharmaceutical industry. A recent investigation into the origins of 156 clinical candidates[5] found that only 1% came from virtual screening; in contrast, over 90% of clinical candidates were derived from patent busting or high throughput screening (HTS). Unfortunately, these sources are increasingly challenged, given the pharmaceutical industry's shift to novel target classes, such as proximity-induced protein degradation[6], protein–protein interactions[7], and RNA targeting[8].

Currently, HTS is the critical tool in drug discovery, providing most novel scaffolds of recent clinical candidates[5,9,10]. These initial starting points crucially shape the course of downstream medicinal chemistry efforts, as most drugs preserve at least 80% of the scaffold of the initially identified lead[11]. Despite these foundational contributions, HTS suffers from practical limitations. Principally, HTS, like all physical experiments, requires that the compounds exist. However, with the advent of synthesis-on-demand libraries, most commercially-available molecules have yet to be synthesized. Still, they can be made and delivered for testing in a matter of weeks[12–14]. These libraries comprise trillions of molecules[14,15] that exemplify millions of otherwise-unavailable scaffolds[12], providing an opportunity to substantially expand the scope and diversity of available chemical space explored in the standard drug discovery process.

Computational approaches unlock this opportunity by reversing the requirement to make molecules before testing them. When computational experiments replace HTS as the primary screen, molecules are tested *before* they are made, and the results from these experiments can inform which molecules are worth synthesizing. Computational experiments further promise to improve upon HTS in terms of cost, speed, need to produce significant quantities of protein[16], effort of miniaturizing assay formats while maintaining experimental integrity[17–19], and reducing false-positive and false-negative rates[16,20–23] including artifacts from aggregation, covalent modification of the target, autofluorescence, or interactions with the reporter rather than the target[20,24,25]. Historical computational techniques such as ligand-based QSAR[26–28], structure-based docking[29,30], and machine learning[31,32] purport to address these limitations of physical screening methods. Unfortunately, these techniques have not replaced HTS; in fact, despite increasing interest in ML, the proportion of drugs discovered with computational techniques has remained steady over the past decades[5,10].

Because there will always be individual targets for which one screening technique can identify more hits than another, the key question governing if computation is ready to be the default hit discovery technique is whether computational screens can identify hits successfully across a broad range of diverse targets. Unfortunately, despite excellent benchmark accuracies[33–35], prospective discovery accuracy remains modest[33,36,37]. For

[1]San Fransisco, CA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉email: izhar@atomwise.com

example, Cerón-Carrasco[38] reported over 700 virtual screens against the SARS-CoV-2 main protease. However, when the author sought to validate the computational predictions via physical experiments, the identified compounds were barely active (800uM). Computational approaches have also been limited by a need for extensive target-specific training data[31,39–41], a requirement for high-quality X-ray crystal structures[42,43], dependence on human adjudication (so-called 'cherry-picking')[12], or a limited domain of applicability[44–48]. Even recent systems have demonstrated utility only in identifying minor variants of known molecules for well-studied proteins with tens of thousands of known binders in their training data[49,50]. Figure 1 exemplifies the striking similarities between recently ML-developed compounds and their preceding published chemical matter. This is particularly concerning, as a myopic focus on well-studied proteins has been identified as a cause of low productivity in pharmaceutical discovery[51].

Nevertheless, we have observed that deep learning approaches are not as limited as these historical examples would imply. Using our AtomNet[52–54] screening system, we have previously reported success in finding novel scaffolds for targets without known ligands[55–57], X-ray crystal structures[56–60], or both[56,57], as well as challenging modulation via protein–protein interaction[59,61] or allosteric binding[60] (see Supplementary Table S1 for examples). However, individual examples do not demonstrate the overall success of such deep learning systems. We therefore report our internal discovery efforts against 22 targets of pharmaceutical interest. We then attempted to further assess the generalizability and robustness of deep learning predictive systems by identifying bioactive molecules for a diverse set of targets. We partnered with 482 academic labs and screening centers, from 257 different academic institutions across 30 countries, through our academic collaboration program, the Artificial Intelligence Molecular Screen (AIMS). This collaboration afforded an opportunity to prospectively evaluate the utility of the AtomNet model as a primary screen across a broad range of diverse, challenging, and realistic targets. In aggregate, we report successes and failures from 318 prospective experiments and evaluate our AtomNet machine-learning technology's ability to serve as a viable alternative to physical HTS campaigns.

Results

We investigated the ability of deep learning-based methods to identify novel bioactive chemotypes by applying the AtomNet model to identify hits for 22 internal targets of pharmaceutical interest. We also explored the breadth of applicability of this approach by attempting to identify drug-like hits in single-dose screens for 296 academic targets, of which 49 were followed up with dose–response experiments, and 21 were further validated by exploring analogs of the initial hits. The average hit rate for our internal projects (6.7%) was comparable to the hit rate for our academic collaborations (7.6%).

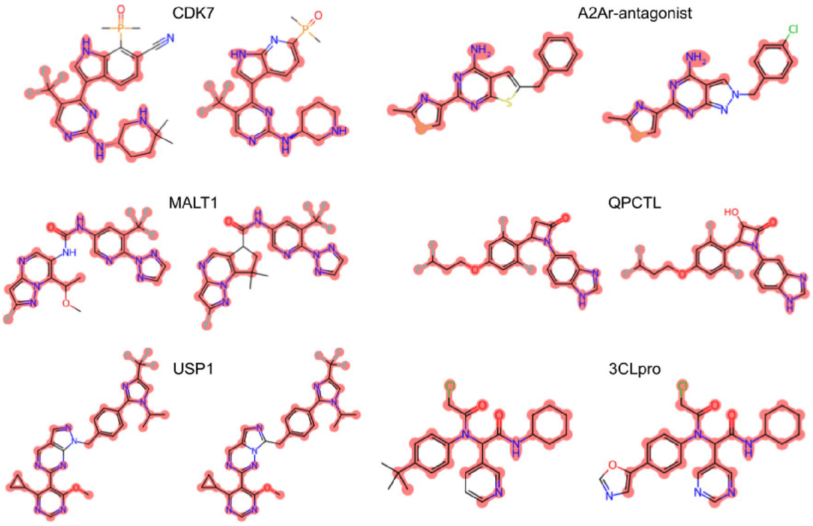


Figure 1. Pairs of representative compounds extracted from AI patents (right) and corresponding prior patents (left) for clinical-stage programs (CDK7[92,93], A2Ar-antagonist[94,95], MALT1[96,97], QPCTL[98,99], USP1[100,101], and 3CLpro[102,103]). The identical atoms between the chemical structures are highlighted in red.

Internal portfolio validation

As part of Atomwise's internal drug discovery efforts, we used the AtomNet model instead of high-throughput or DNA-encoded library (DEL) screening. We screened a 16-billion synthesis-on-demand chemical space[62], which is several thousand times larger than HTS libraries and even exceeds the size of most DELs without suffering limitations of DNA-compatible chemistry[16,23]. Each screen requires over 40,000 CPUs, 3,500 GPUs, 150 TB of main memory, and 55 TB of data transfers. We describe the protocol in detail in the Methods section; briefly, we computationally scored each catalog compound after removing molecules that were prone to interfere with the assays or were too similar to known binders of the target or its homologs. The neural network analyzes and scores the 3D coordinates of each generated protein–ligand co-complex, producing a list of ligands ranked by their predicted binding probability. Our workflow then clusters the top-ranked molecules to ensure diversity and algorithmically selects the highest-scoring exemplars from each cluster. At no point are compounds manually cherry-picked. The molecules were synthesized at Enamine (https://enamine.net) and quality controlled by LC–MS to purity > 90%, in agreement with HTS standards[63]. Hits were further validated using NMR. We then physically tested, on average, 440 compounds per target at reputable contract research organizations (CROs), while attempting to mitigate assay interferences such as aggregation and oxidation with standard additives (*e.g.*, Tween-20, Triton-X 100, and dithiothreitol (DTT)). We describe the assay protocols in detail in the Supplementary Data S1.

We describe the results of the 22 experiments in Table 1. In 91% of the experiments, we identified single-dose (SD) hits that were reconfirmed in dose–response (DR) experiments. The average target DR hit rate was 6.7% compared to 8.8% from the SD screens. Only 16 of the 22 projects were structurally enabled with X-ray crystallography; one used a cryo-EM structure, while five used homology models with an average sequence identity of 42% to their template protein. The DR hit rate for the cryo-EM project was 10.56%, while the average hit rate for the homology models was a similar 10.8%.

We then advanced 14 projects with at least one dose-responsive scaffold to a round of analog expansion. We found new bioactive analogs in the SD screen for all projects, with an average hit rate of 29.8%. Further validation with DR resulted in an average hit rate of 26% per project, which compares favorably with typical HTS hit rates ranging from 0.151 to 0.001%[64,65]. We note that the size and chemical diversity within and between physical[66] and virtual[14] HTS libraries prevent an explicit evaluation of the methods over the same chemical space. The most potent analogs ranged from single-digit nanomolar, against a kinase, to double-digit micromolar, against a transcription factor (Supplementary Table S2). Additionally, we present two internal studies in detail. For Large Tumor Suppressor Kinase 1 (LATS1), we identified potent compounds despite the lack of a crystal structure or known active compounds. For ATP-driven chaperone Valosin Containing Protein (VCP) we identified novel allosteric and orthosteric modulators.

Gene name	# of compounds tested	SD hit rate (%)	DR hit rate (%)	Potency range (IC50/Ki, uM)	# of analog tested	SD analog hit rate (%)	DR analog hit rate (%)	Analog potency range (IC50/Ki, uM)
ASAH1	376	10.64	7.71	0.3–102	–	–	–	–
AXL	597	12.06	8.21	0.181–71	3200	35.59	33.56	0.079–86
BCL2	422	3.08	0.00	–	–	–	–	–
CBLB	422	1.66	0.00	–	–	–	–	–
CDK5	786	10.69	10.43	0.049–79	587	47.53	43.61	0.43–76
CDK7	786	10.69	10.56	0.099–60	735	28.44	27.35	0.191–10
GFPT1	384	6.51	2.34	31–86	734	24.93	24.11	1–194
KCNT1	416	9.62	7.69	1.1–30	–	–	–	–
KDM6A	356	3.93	1.12	24–58	–	–	–	–
LATS1	418	18.18	17.94	0.077–82	841	51.72	45.78	0.034–98
MC2R	208	11.54	9.62	16–68	419	39.38	38.42	2.4–97
MDM4	422	2.37	0.47	5.9–29.8	192	18.23	18.23	4.4–90
NT5E	335	1.49	0.30	176	221	9.95	1.81	8.3–65
PARG	334	7.78	7.78	15–250	–	–	–	–
PARP14	576	5.38	2.95	3–96	616	26.46	26.30	0.2–95
POLQ	330	11.82	11.52	1.2–49	559	11.27	8.77	1.5–42
PPARA	422	4.03	0.24	131	211	14.22	3.79	59–95
PPM1D	530	11.89	6.98	4.5–98	–	–	–	–
PRMT5	422	4.03	0.95	7.2–79	415	7.95	5.54	19–114
PRODH2	542	2.77	1.11	15–84	–	–	–	–
TYK2	189	38.10	34.39	0.016–9	457	71.33	60.39	0.006–10
VCP	416	4.81	4.81	2.4–64	738	–	–	–

Table 1. Results from 22 Atomwise internal programs. SD and DR denote single-dose and dose–response, respectively.

Academic validation

In addition to our internal discovery efforts, we performed virtual screens for 296 targets, comprising more than 20 billion individual neural network scores of generated protein–ligand co-complexes. We purchased, on average, 85 off-the-shelf commercially available compounds, quality controlled by NMR and LC–MS to > 90% purity[63], and plated in a single 96-well plate. The compounds were then physically screened for activity against the target of interest in single-dose assays (see Supplemental Data S1 for assay protocols). As with HTS primary screens, additional characterization studies are required to validate the initially identified hits so, in 49 projects, we performed dose–response studies and analog expansion. We present a summary of our results in Supplementary Table S3.

Figure 2 illustrates the distributions of projects across therapeutic areas, protein families, and assay types. Every major therapeutic area is represented, with the most frequent area being oncology, comprising 35% of projects, followed by infectious diseases and neurology, comprising 27% and 9% of projects, respectively. Breaking down the projects by protein families reveals that all major enzyme classes are represented, with enzymes comprising 59% of the targets and membrane proteins such as GPCR, transporters, and ion channels, representing 12% of the targets. Working on a large and diverse set of therapeutic targets requires a heterogeneous collection of biological assays; 20% of the assays measured direct binding, whereas 56% and 20% were functional and phenotypic.

In 215 projects, we identified at least one bioactive compound for the target in a biochemical or cell-based assay. This 73% success rate substantially improves over the ~50% success rate for HTS[21,67]. On average, we screened 85 compounds per project and discovered 4.6 active hits, with an average hit rate of 5.5%. For the subset of targets where we found any hits, the average was 6.4 hits per project. Thus, we achieved an average hit rate of 7.6%, which again compares favorably with typical HTS hit rates. See Supplementary Material S1 for all assay definitions and conditions. Supplementary Table S4 shows a representative bioactive compound from each of the 215 successful projects, and Supplementary Fig. S2 shows that the physicochemical properties of the identified hits are largely druglike and Lipinski-compliant.

The AtomNet technology robustly identified active molecules, even for targets that lacked prior on-target bioactivity data. This ability to identify hits for previously undrugged targets is critical if machine learning-based approaches are to replace HTS as the default primary screening approach. For 207 out of the 296 targets (70%), the training data available for AtomNet models lacked a single active molecule for that target or any closely related protein (i.e., proteins with sequence identity greater than 70%). We interpret this as evidence of the ability of properly-architected machine learning systems to extrapolate to novel biological space. Figure 3A illustrates the hit rate versus the number of training examples available to our model. Although previous computational

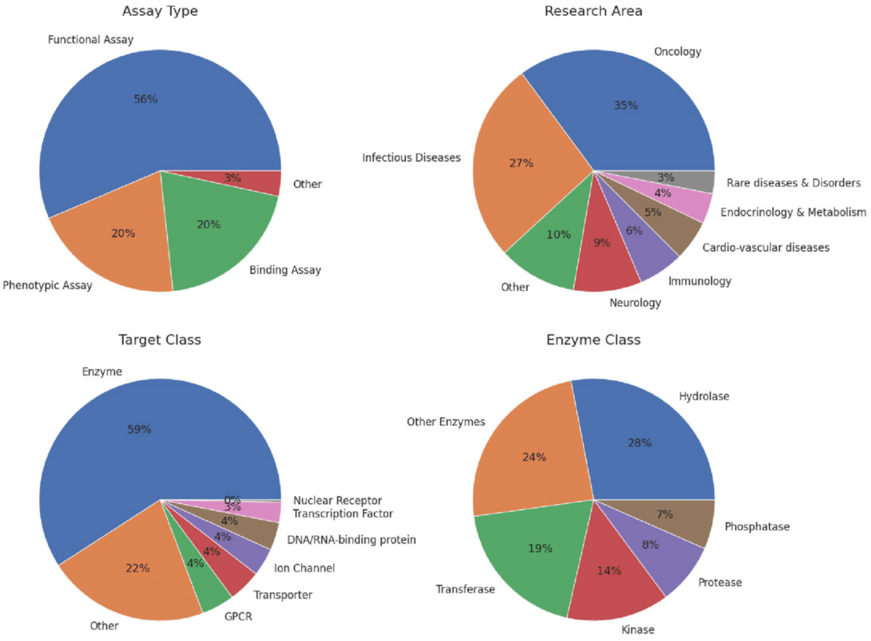


Figure 2. The distributions of 296 AIMS projects across assay types used in the primary screen, research areas, target classes, and further breakdown to enzyme classes when applicable.

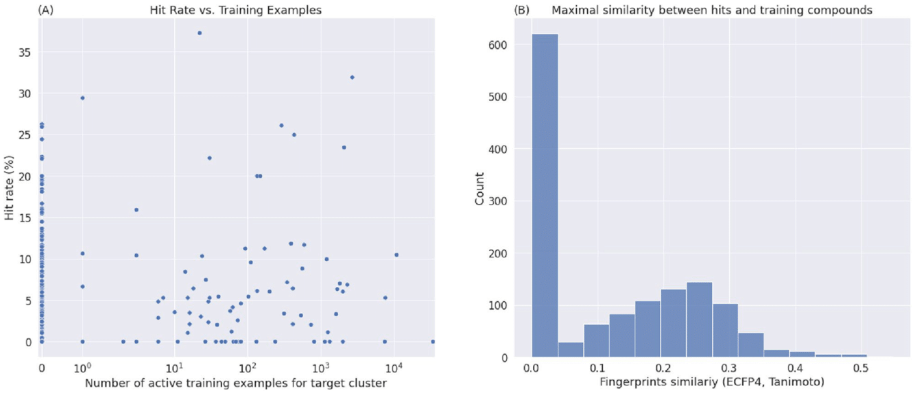


Figure 3. (**A**) An illustration of the hit rate versus the number of training examples available to our model. Each point represents a project, with the x-axis denoting the number of active molecules in our training for the target protein or homologs and the y-axis denoting the hit rate of the project (the percentage of molecules tested in the project that were active). The model shows no dependence on the availability of on-target training examples. For 70% of the targets, the AtomNet model training data lacked any active molecules for that target or any similar targets with greater than 70% sequence identity, yet the model achieved a hit rate of 5.3% compared to 6.1% when on-target data was available. (**B**) The distribution of similarities between hits and their most-similar bioactive compounds in our training data. Our screening protocol ensures that the compounds subjected to physical testing are not similar to known active compounds or close homologs (< 0.5 Tanimoto similarity using ECFP4, 1024 bits). Because 70% of the AIMS targets had no annotated bioactivities in our training dataset, hits identified in these projects have a similarity value of zero.

approaches typically require thousands of on-target training examples[31,39,42], the lack of correlation between training examples and hit rate ($R^2 = 0.0021$, p-value = 0.43) shows that our ML algorithm is agnostic to the availability of such data. We achieved an average success rate of 75% and hit rates of 5.3% when no training data was available, comparable to the 67% and 6.1% success and hit rates achieved when binding data was available in the training set. Interestingly, we also do not see a significant increase in hit rate attributable to the proportion of binding data available for a target ($R^2 = 0.008$, p-value = 0.39). This reflects the robustness of the screening protocol and the chemical dissimilarity of scaffolds identified by AtomNet models to previously known bioactive compounds.

Next, we assessed the ability of the AtomNet models to identify novel scaffolds. This is a critical capability for primary screens, as follow-up assays tend to work within the chemical space uncovered in the initial screen. The task of novel scaffold identification appears in two distinct scenarios: (1) when no scaffold is known for the target and we wish to identify the first scaffold, and (2) when some scaffolds are known but we wish to identify dissimilar scaffolds because novel chemical matter can yield improved selectivity, toxicity, pharmacokinetics, or patentability. Performance of AtomNet models for the first scenario, when no scaffolds for the target existed in the AtomNet model training data, was evaluated on 70% of the targets, where the training data contained no active molecules for the target or its homologs (vide supra). We achieved an average hit rate of 5.3% for targets with no training data. For the second scenario, we analyzed the similarity of the identified hits to known bioactive compounds in our training data (Fig. 3B). Our screening protocol ensures that the compounds subjected to physical testing are not similar to known active compounds or close homologs (< 0.5 Tanimoto similarity using ECFP4[68], 1024 bits). We interpret this as evidence of the ability of properly-architected machine learning systems to extrapolate to novel chemical space as well. For cases where training data was available (i.e., the Tanimoto similarity is above zero), the similarity distribution is close to the one expected by random compound pairs[69]. The novelty of the small-molecule structures is striking because target-specific machine-learning algorithms tend to uncover highly similar analogs for known bioactive molecules[50,70,71]. The superior performance of the AtomNet model is expected, considering the bias-variance tradeoff[72] in machine learning algorithms. Because the AtomNet convolutional neural network is a global model, concurrently trained on millions of bioactivities, hundreds of thousands of small molecules, and thousands of protein binding sites, it can reduce both bias and variance of the model compared to target-specific ones[33]. Specifically, our global model can benefit from multiple levels of information captured in the structures of the small molecules, the sequences of the target proteins, and the three-dimensional interactions between the two.

AtomNet also successfully identified active molecules when there was no X-ray crystal structure of the receptor. Figure 4A compares the hit rates obtained with 3-dimensional crystal structures, cryo-EM, and homology modeling. We did not attempt to select targets based on the similarity to the template but rather used the best template available. We observe no substantial difference in success rate between the three, in contrast to the common challenges in using homology models or low-precision structures for structure-based discovery[42,43,73]. We achieved average hit rates of 5.6%, 5.5%, and 5.1% for crystal structures, cryo-EM, and homology modeling. We

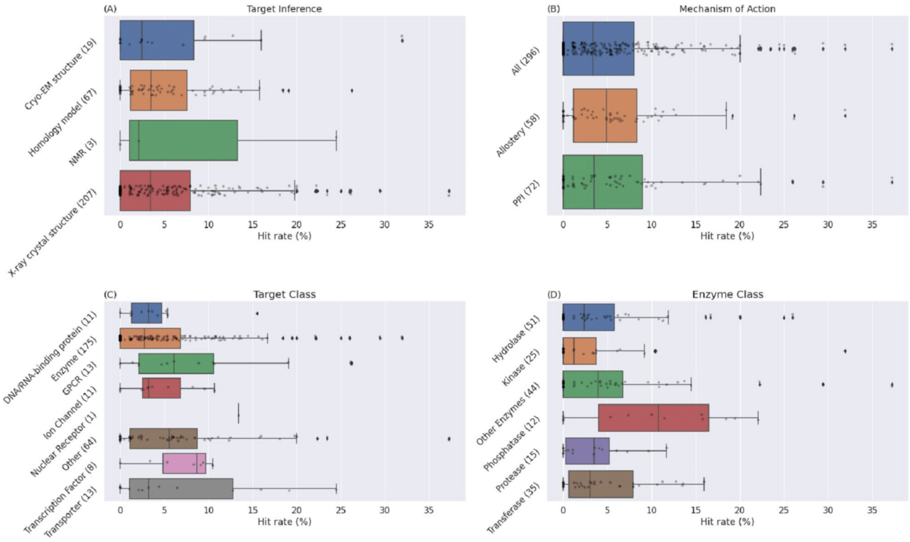


Figure 4. Hit rates obtained for the 296 AIMS projects. (**A**) A comparison of hit rates using X-ray crystallography, NMR, Cryo-EM, and homology for modeling the structure of the proteins. Each point represents a project with the x-axis denoting the hit rate of the project (the percentage of molecules tested in the project that were active). The number of projects of each type is given in parentheses. We observed no substantial difference in success rate between the physical and the computationally inferred models. We achieved average hit rates of 5.6%, 5.5%, and 5.1% for crystal structures, cryo-EM, and homology modeling, respectively. The number of projects using NMR structures is too small to make statistically-robust claims. (**B**) A comparison of hit rates observed for traditionally challenging target classes such as protein–protein interactions (PPI) and allosteric binding. Of the 296 projects, 72 targeted PPIs and 58 allosteric binding sites. The average hit rates were 6.4% and 5.8% for PPIs and allosteric binding, respectively. (**C**) Comparison of hit rates observed for different target classes and (**D**) enzyme classes. No protein or enzyme class falls outside the domain of applicability of the algorithm.

also successfully identified active compounds in projects with NMR structures, but the number of such targets is too small to make statistically-robust claims.

An interesting demonstration of the robustness of the AtomNet model to low data and poorly characterized protein structure is its ability to identify novel hits for traditionally challenging target classes such as protein–protein interaction (PPI) sites and allosteric binding sites (Fig. 3B). Of the 296 projects, 72 targeted PPIs and 58 allosteric binding sites. We identified hits for 53 (74%) PPI sites and 46 (79%) allosteric sites, with 13 projects representing allosteric sites at PPI interfaces. The average hit rate was 6.4% and 5.8% for PPIs and allosteric binding sites, respectively. The algorithm's success in these target classes, which often suffer from poorly characterized binding sites and a lack of bioactivity training data, is not surprising because Fig. 2A shows that our model is largely not dependent on the availability of on-target training data.

Finally, we investigated whether the algorithm exhibits domain of applicability limitations regarding different protein classes. Figures 4C and 3D illustrate the hit rate observed for each protein and enzyme class. No protein or enzyme class falls outside the domain of applicability of the algorithm, demonstrating that machine learning-based approaches are well-suited as a default technology for new scaffold identification. The hit rate for nuclear receptors is an outlier, with seemingly better accuracy than other classes, but a single data point is not statistically meaningful.

Dose–response validation studies

We performed additional validation studies for 49 AIMS projects with at least one reported hit. The objective of the validation studies was to establish dose–response (DR) relationships for the single-dose (SD) hits. We describe the protocol of the DR experiments in the Methods section. Briefly, we performed dose–response measurements for the reported hits from the single-dose primary screens. DR was determined using the same assay and screening protocol as the single-dose screens, at the same lab, and with the same personnel. Full dose response curves were obtained in most cases, however in some instances a full curve was not obtained, or concentration dependent activity was qualitatively determined by testing at concentrations other than that for the primary screen. The distribution of assay types and target classes for the projects selected for DR validation also was similar to that of the AIMS projects (Supplementary Fig. S3).

We describe the results of the DR experiments in Supplementary Table S5. In 84% of the experiments, we validated at least one SD hit and got a DR readout. The median activity for the total of 144 DR measurements was 15.4 µM (which compares favorably with HTS[25,74]), of which 13% showed sub-µM potency. Overall, we achieved an average of 2.8 hits per validation study, resulting in a hit rate of 51%. The false positive rate of 49% observed in these experiments is favorably compared to HTS' which can be as high as 95%[20,75]. This difference in false positive rates may stem from the comparative ease and robustness of the low-throughput assay format we employed versus high-throughput assay. Representative dose–response curves for each of the 49 projects are shown in Supplementary Table S6.

Analog validation studies

For a subset of 21 projects, we further validated hits with DR activity by testing analogs of the active compounds. In those cases, we used the AtomNet platform to search a purchasable space for additional bioactive compounds chemically analogous to the SD hits. We selected up to 35 additional compounds for testing, including the active compounds from the SD screens.

We describe the results of the analoging experiments in Supplementary Table S7. We identified additional analogs with DR readouts for 16 projects (76%). The median DR activity of the 154 validated analogs was 7.4 µM compared to the median of 15.4 µM of the parent compound (Supplementary Fig. S4).

Methods

Screening protocols

AIMS screening protocol

We began by evaluating screening libraries of millions of catalog compounds from commercial vendors MCule (10 M)[76] and Enamine in-stock (2.5 M)[77]. We then selected a drug-like subset via algorithmic filtering by applying Eli Lilly medicinal chemistry filters[78] and removing likely false positives, such as aggregators, autofluorescers, and PAINS[79] (see Fig. 2 for the distributions of drug-like properties of the SD hits). The resulting library was virtually screened against the target of interest, removing any molecules with greater than 0.5 Tanimoto similarity in ECFP4 space to any known binders of the target and its homologs within 70% sequence identity. For kinase targets, we extend the exclusion to the whole kinome. The binding site was defined using co-complexes, mutagenesis studies, co-complexes of homologs, or by identifying potential sites using ICM Pocket Finder[80] or Fpocket[81]. Some were orthosteric, while others were allosteric, or as yet unestablished biological functions. In 64 cases, we built homology models using the closest sequence, with an average sequence similarity of 54%. We clustered the top 30,000 molecules using the Butina[82] algorithm with a Tanimoto similarity cutoff of 0.35 in ECFP4 space, selecting the highest-scoring exemplars. Additional computed physico-chemical property filters were applied as needed. At no point were compounds cherry-picked. We purchased, on average, 85 compounds, quality controlled by LC–MS to > 90% purity, generally dispensed as 10 mM DMSO stocks plated in a single 96-well plate. In addition, two vials of DMSO-only negative controls were included before scrambling the compound locations on the plate, by the supplier, for blinded experimental testing. To further control for potential artifacts, we removed compounds that showed measurable activity toward more than one target from the analysis.

Dose–response and analoging validation screening protocol

We considered advancing AIMS projects to additional validation studies based on the ability to reorder at least some of the initial SD hits, the availability of chemical analogs in the screening library to the initial hits, the capability to perform dose–response experiments, and the ability of the collaborators to perform additional screens and return results promptly.

We performed two sets of experiments: DR validation of the SD hits from AIMS and analoging with DR readouts. We performed DR measurements using the same assays and protocols as SD.

We performed an analoging round by identifying, for each AIMS hit, its 1000 nearest neighbors from the Mcule library[76], using molecular fingerprints similarity[68]. We augmented the set with additional analogs using substructure[83] or FTrees[84] searches, if needed. We used an AtomNet regression model, trained to predict quantitative bioactivities (e.g., IC50 or Ki), to score and rank the analogs. A set of 20—35 compounds from the analogs space of an initial hit were then obtained based on similarity and top scores from the AtomNet model for testing.

Internal portfolio screening protocol

We followed a protocol similar to the AIMS screen with a few deviations. First, we used the Enamine REAL library of over 16 billion compounds[62]. Second, we used an ensemble of six AtomNet models for the screens. Last, on average, we selected a set of 440 compounds for testing.

The analoging protocol is similar to the AIMS validation studies, with the following deviations. First, we used the Enamine REAL library for analog search. Second, we selected an average of 676 analogs per project. Third, the analog search protocol was more complex, pulling nearest neighbors based on maximum common substructure and graph edit distance in addition to the ECFP4-based one.

AtomNet® model architecture

We previously published in detail[52,53,55,58,59,61,85,86] during the course of the AIMS program, and we described the most recent version of the AtomNet model architecture in detail elsewhere[53]. We provide a brief description below.

The AtomNet model is a Graph Convolution Network architecture with atoms represented as vertices and pair-wise, distance-dependent, edges representing atom proximities. The input is a graph network of features characterizing the atom types and topologies of an ensemble of protein–ligand complexes. Receptor atoms more than 7 Å away from any ligand atom are excluded from the complexes, and each node in the graph is associated with a feature vector representing the atom type using Sybyl typing[87].

The network has five graph convolutional blocks. In the first two graph convolution blocks, all ligand and receptor atoms 5 Å apart from each other are considered, and 64 filters per block are used. In the third block, the cutoff radius and filters are increased to 7 Å and 128, respectively. Only ligand features in the last two blocks are considered without changing the threshold cutoff or the number of filters. Finally, the sum-pool of the ligand-only layer creates a 3-task layer on top of the network. That multi-task layer predicts three endpoints: bioactivity, pose quality, and a physics-based docking score[88].

We trained an ensemble of 6 models, splitting the training data into sixfold cross-validation sets based on a protein sequence similarity cutoff of 70%. Then, each model in the ensemble was trained on a different fold for 10 epochs, using the ADAM optimizer[89] with a learning rate of 0.001, and targets were sampled with replacement, proportional to the number of active compounds associated with that target.

Data

All data generated or analyzed during this study are included in this published article (and its supplementary information S1 files). Boxplots illustrations show the quartiles (Q1 and Q3) of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be “outliers” (1.5 × of the inter-quartile range, as implemented in the Seaborn and Matplotlib toolboxes[90,91]).

Conclusion

HTS is the most widely-used tool for hit discovery for new targets. Unfortunately, all physical screening methods share the critical limitation that a molecule must exist to be screened. Computational methods enable a fundamental shift to a test-then-make paradigm. In this work, we report on 318 projects (22 internal projects and 296 collaborations) where we used the AtomNet platform as the primary screening tool coupled with low-throughput physical screens as validation. The AtomNet technology can identify bioactive scaffolds across a wide range of proteins, even without known binders, X-ray structures, or manual cherry-picking of compounds. Our empirical results suggest that machine learning approaches have reached a computational accuracy that can replace HTS as the first step of small-molecule drug discovery.

Data availability

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Received: 15 September 2023; Accepted: 15 February 2024
Published online: 02 April 2024

References

1. Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082 (1992).
2. Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **1**, 882–894 (2002).
3. Walters, W. P., Stahl, M. T. & Murcko, M. A. Virtual screening—an overview. *Drug Discov. Today* **3**, 160–178 (1998).
4. Ring, C. S. *et al.* Structure-based inhibitor design by using protein models for the development of antiparasitic agents. *Proc. Natl. Acad. Sci. USA*. **90**, 3583–3587 (1993).
5. Brown, D. G. An analysis of successful hit-to-clinical candidate pairs. *J. Med. Chem.* https://doi.org/10.1021/acs.jmedchem.3c00521 (2023).
6. Békés, M., Langley, D. R. & Crews, C. M. PROTAC targeted protein degraders: The past is prologue. *Nat. Rev. Drug Discov.* **21**, 181–200 (2022).
7. Lu, H. *et al.* Recent advances in the development of protein–protein interactions modulators: Mechanisms and clinical trials. *Signal Transduct. Target. Ther.* **5**, 1–23 (2020).
8. Childs-Disney, J. L. *et al.* Targeting RNA structures with small molecules. *Nat. Rev. Drug Discov.* **21**, 736–762 (2022).
9. Brown, D. G. & Boström, J. Where do recent small molecule clinical development candidates come from?. *J. Med. Chem.* **61**, 9442–9468 (2018).
10. Dragovich, P. S., Haap, W., Mulvihill, M. M., Plancher, J.-M. & Stepan, A. F. Small-molecule lead-finding trends across the roche and genentech research organizations. *J. Med. Chem.* **65**, 3606–3615 (2022).
11. Perola, E. An analysis of the binding efficiencies of drugs and their leads in successful drug discovery programs. *J. Med. Chem.* **53**, 2986–2997 (2010).
12. Lyu, J. *et al.* Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224 (2019).
13. Sadybekov, A. A. *et al.* Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452–459 (2022).
14. Bellmann, L., Penner, P., Gastreich, M. & Rarey, M. Comparison of combinatorial fragment spaces and its application to ultralarge make-on-demand compound catalogs. *J. Chem. Inf. Model.* **62**, 553–566 (2022).
15. Neumann, A., Marrison, L. & Klein, R. Relevance of the trillion-sized chemical space “explore” as a source for drug discovery. *ACS Med. Chem. Lett.* **14**, 466–472 (2023).
16. Sunkari, Y. K., Siripuram, V. K., Nguyen, T.-L. & Flajolet, M. High-power screening (HPS) empowered by DNA-encoded libraries. *Trends Pharmacol. Sci.* **43**, 4–15 (2022).
17. Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167–175 (2006).
18. Iversen, P. W., Eastwood, B. J., Sittampalam, G. S. & Cox, K. L. A comparison of assay performance measures in screening assays: Signal window, Z' factor, and assay variability ratio. *J. Biomol. Screen.* **11**, 247–252 (2006).
19. Zhang, J.-H., Chung, T. D. Y. & Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **4**, 67–73 (1999).

20. Jadhav, A. *et al.* Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. *J. Med. Chem.* **53**, 37–51 (2010).
21. Fox, S. *et al.* High-throughput screening: Update on practices and success. *J. Biomol. Screen.* **11**, 864–869 (2006).
22. Owen, S. C., Doak, A. K., Wassam, P., Shoichet, M. S. & Shoichet, B. K. Colloidal aggregation affects the efficacy of anticancer drugs in cell culture. *ACS Chem. Biol.* **7**, 1429–1435 (2012).
23. Rössler, S. L., Grob, N. M., Buchwald, S. L. & Pentelute, B. L. Abiotic peptides as carriers of information for the encoding of small-molecule library synthesis. *Science* **379**, 939–945 (2023).
24. McGovern, S. L., Caselli, E., Grigorieff, N. & Shoichet, B. K. A Common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **45**, 1712–1722 (2002).
25. Feng, B. Y., Shelat, A., Doman, T. N., Guy, R. K. & Shoichet, B. K. High-throughput assays for promiscuous inhibitors. *Nat. Chem. Biol.* **1**, 146–148 (2005).
26. Martin, E. J., Polyakov, V. R., Tian, L. & Perez, R. C. Profile-QSAR 2.0: Kinase virtual screening accuracy comparable to four-concentration IC50s for realistically novel compounds. *J. Chem. Inf. Model.* **57**, 2077–2088 (2017).
27. Keiser, M. J. *et al.* Predicting new molecular targets for known drugs. *Nature* **462**, 175–181 (2009).
28. Svetnik, V. *et al.* Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **43**, 1947–1958 (2003).
29. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
30. Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **432**, 862–865 (2004).
31. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **55**, 263–274 (2015).
32. Sheridan, R. P. *et al.* Machine Learning and Deep Learning Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of QSAR models?. *J. Chem. Inf. Model.* https://doi.org/10.1021/acs.jcim.9b01067 (2020).
33. Wallach, I. & Heifets, A. Most ligand-based classification benchmarks reward memorization rather than generalization. *J. Chem. Inf. Model.* **58**, 916–932 (2018).
34. Chen, L. *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLOS ONE* **14**, e0220113 (2019).
35. Chuang, K. V. & Keiser, M. J. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362**, eaat8603 (2018).
36. Gaieb, Z. *et al.* D3R Grand Challenge 3: Blind prediction of protein–ligand poses and affinity rankings. *J. Comput. Aided Mol. Des.* **33**, 1–18 (2019).
37. Gabel, J., Desaphy, J. & Rognan, D. Beware of machine learning-based scoring functions on the danger of developing black boxes. *J. Chem. Inf. Model.* **54**, 2807–2815 (2014).
38. Cerón-Carrasco, J. P. When virtual screening yields inactive drugs: dealing with false theoretical friends. *ChemMedChem* **17**, e202200278 (2022).
39. McCloskey, K. *et al.* Machine learning on DNA-encoded libraries: A new paradigm for hit-finding. *J. Med. Chem.* **63**, 8857–8866 (2020).
40. Wenzel, J., Matter, H. & Schmidt, F. Predictive multitask deep neural network models for ADME-Tox properties: Learning from large data sets. *J. Chem. Inf. Model.* **59**, 1253–1268 (2019).
41. Feinberg, E. N. *et al.* PotentialNet for molecular property prediction. *ACS Cent. Sci.* **4**, 1520–1530 (2018).
42. Schindler, C. E. M. *et al.* Large-scale assessment of binding free energy calculations in active drug discovery projects. *J. Chem. Inf. Model.* **60**, 5457–5474 (2020).
43. Bordogna, A., Pandini, A. & Bonati, L. Predicting the accuracy of protein–ligand docking on homology models. *J. Comput. Chem.* **32**, 81–98 (2011).
44. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688-702.e13 (2020).
45. Melo, M. C. R., Maasch, J. R. M. A. & de la Fuente-Nunez, C. Accelerating antibiotic discovery through artificial intelligence. *Commun. Biol.* **4**, 1–13 (2021).
46. Skinnider, M. A. *et al.* A deep generative model enables automated structure elucidation of novel psychoactive substances. *Nat. Mach. Intell.* **3**, 973–984 (2021).
47. Muegge, I. & Oloff, S. Advances in virtual screening. *Drug Discov. Today Technol.* **3**, 405–411 (2006).
48. N. Muratov, E. *et al.* QSAR without borders. *Chem. Soc. Rev.* **49**, 3525–3564 (2020).
49. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
50. Walters, W. P. & Murcko, M. Assessing the impact of generative AI on medicinal chemistry. *Nat. Biotechnol.* **38**, 143–145 (2020).
51. Scannell, J. W., Blanckley, A., Boldon, H. & Warrington, B. Diagnosing the decline in pharmaceutical R&D efficiency. *Nat. Rev. Drug Discov.* **11**, 191 (2012).
52. Wallach, I., Dzamba, M. & Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. *ArXiv Prepr. ArXiv151002855* 1–11 (2015).
53. Gniewek, P., Worley, B., Stafford, K., van den Bedem, H. & Anderson, B. *Learning physics confers pose-sensitivity in structure-based virtual screening.* https://doi.org/10.48550/arXiv.2110.15459 (2021).
54. Stafford, K. A., Anderson, B. M., Sorenson, J. & van den Bedem, H. AtomNet PoseRanker: Enriching ligand pose quality for dynamic proteins in virtual high-throughput screens. *J. Chem. Inf. Model.* **62**, 1178–1189 (2022).
55. Hsieh, C.-H. *et al.* Miro1 marks parkinson's disease subset and miro1 reducer rescues neuron loss in Parkinson's models. *Cell Metab.* **30**, 1131-1140.e7 (2019).
56. Reidenbach, A. G. *et al.* Multimodal small-molecule screening for human prion protein binders. *J. Biol. Chem.* **295**, 13516–13531 (2020).
57. Bon, C. *et al.* Discovery of novel trace amine-associated receptor 5 (TAAR5) antagonists using a deep convolutional neural network. *Int. J. Mol. Sci.* **23**, 3127 (2022).
58. Stecula, A., Hussain, M. S. & Viola, R. E. Discovery of novel inhibitors of a critical brain enzyme using a homology model and a deep convolutional neural network. *J. Med. Chem.* **63**, 8867–8875 (2020).
59. Su, S. *et al.* SPOP and OTUD7A Control EWS–FLI1 protein stability to govern ewing sarcoma growth. *Adv. Sci.* **8**, 2004846 (2021).
60. Pedicone, C. *et al.* Discovery of a novel SHIP1 agonist that promotes degradation of lipid-laden phagocytic cargo by microglia. *iScience* **25**, 104170 (2022).
61. Huang, C. *et al.* Small molecules block the interaction between porcine reproductive and respiratory syndrome virus and CD163 receptor and the infection of pig cells. *Virol. J.* **17**, 116 (2020).
62. Grygorenko, O. O. *et al.* Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
63. Dandapani, S., Rosse, G., Southall, N., Salvino, J. M. & Thomas, C. J. Selecting, acquiring, and using small molecule libraries for high-throughput screening. *Curr. Protoc. Chem. Biol.* **4**, 177–191 (2012).
64. Schuffenhauer, A. *et al.* Library design for fragment based screening. *Curr. Top. Med. Chem.* **5**, 751–762 (2005).

65. Jacoby, E. *et al.* Key aspects of the novartis compound collection enhancement project for the compilation of a comprehensive Chemogenomics drug discovery screening collection. *Curr. Top. Med. Chem.* **5**, 397–411 (2005).
66. Petrova, T., Chuprina, A., Parkesh, R. & Pushechnikov, A. Structural enrichment of HTS compounds from available commercial libraries. *MedChemComm* **3**, 571–579 (2012).
67. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**, 188–195 (2011).
68. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
69. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminformatics* **5**, 26 (2013).
70. Ren, F. *et al.* AlphaFold accelerates artificial intelligence powered drug discovery: Efficient discovery of a novel cyclin-dependent kinase 20 (CDK20) Small Molecule Inhibitor (2022).
71. Assessing structural novelty of the first AI-designed drug candidates to go into human clinical trials. *CAS* https://www.cas.org/resources/blog/ai-drug-candidates.
72. Kohavi, R. & Wolpert, D. Bias plus variance decomposition for zero-one loss functions. in *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning* 275–283 (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1996).
73. Ferrara, P. & Jacoby, E. Evaluation of the utility of homology models in high throughput docking. *J. Mol. Model.* **13**, 897–905 (2007).
74. Walters, W. P. & Namchuk, M. Designing screens: How to make your hits a hit. *Nat. Rev. Drug Discov.* **2**, 259–266 (2003).
75. Inglese, J. *et al.* High-throughput screening assays for the identification of chemical probes. *Nat. Chem. Biol.* **3**, 466–479 (2007).
76. mcule database. https://mcule.com/database/.
77. Screening Collections - Enamine. https://enamine.net/compound-collections/screening-collection.
78. Bruns, R. F. & Watson, I. A. Rules for identifying potentially reactive or promiscuous compounds. *J. Med. Chem.* **55**, 9763–9772 (2012).
79. Baell, J. B. & Holloway, G. A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J. Med. Chem.* **53**, 2719–2740 (2010).
80. Abagyan, R. & Kufareva, I. The flexible pocketome engine for structural chemogenomics. *Methods Mol. Biol. Clifton NJ* **575**, 249–279 (2009).
81. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics* **10**, 168 (2009).
82. Butina, D. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
83. *RDKit: Open-Source Cheminformatics.*
84. Rarey, M. & Dixon, J. S. Feature trees: A new molecular similarity measure based on tree matching. *J. Comput. Aided Mol. Des.* **12**, 471–490 (1998).
85. Stafford, K., Anderson, B. M., Sorenson, J. & van den Bedem, H. *AtomNet PoseRanker: Enriching Ligand Pose Quality for Dynamic Proteins in Virtual High Throughput Screens.* https://doi.org/10.26434/chemrxiv-2021-t6xkj (2021).
86. Schroedl, S. Current methods and challenges for deep learning in drug discovery. *Drug Discov. Today Technol.* **32–33**, 9–17 (2019).
87. Bender, A., Mussa, H. Y., Glen, R. C. & Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a Naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **44**, 170–178 (2004).
88. Trott, O. & Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
89. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *ArXiv14126980 Cs* (2017).
90. Waskom, M. L. seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).
91. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
92. Marineau, J. J. *et al.* Discovery of SY-5609: A selective, noncovalent inhibitor of CDK7. *J. Med. Chem.* **65**, 1458–1480 (2022).
93. Gu, X., BAI, H., Barbeau, O. R. & Besnard, J. Aromatic heterocyclic compound, and pharmaceutical composition and application thereof. (2022).
94. Barbay, J. K., Chakravarty, D., Leonard, K., Shook, B. C. & Wang, A. Phenyl and heteroaryl substituted thieno[2,3-d]Pyrimidines and their use as adenosine A2a receptor antagonists (2010).
95. Bell, A. S., Schreyer, A. M. & Versluys, S. Pyrazolopyrimidine compounds as adenosine receptor antagonists (2019).
96. Soldermann, C. P. *et al.* Pyrazolo pyrimidine derivatives and their use as MALT1 inhbitors (2019).
97. Feng, S. *et al.* Tricyclic compounds useful in the treatment of cancer, autoimmune and inflammatory disorders (2023).
98. Heiser, U. & Sommer, R. Inhibitors of glutaminyl cyclase (2020).
99. Cheng, X., Liu, Y., Qin, L., Ren, F. & Wu, J. Beta-lactam derivatives for the treatment of diseases (2023).
100. Wylie, A. A. *et al.* Therapeutic combinations comprising ubiquitin-specific-processing protease 1 (usp1) inhibitors and poly (adp-ribose) polymerase (parp) inhibitors (2021).
101. Wu, J., Qin, L. & Liu, J. Small molecule inhibitors of ubiquitin specific protease 1 (usp1) and uses thereof 2023).
102. Stille, J. *et al.* Design, Synthesis and Biological Evaluation of Novel SARS-CoV-2 3CLpro Covalent Inhibitors. https://doi.org/10.26434/chemrxiv.13087742.v1 (2020).
103. Zavoronkovs, A., Ivanenkov, Y. A. & Zagribelnyy, B. Sars-cov-2 inhibitors having covalent modifications for treating coronavirus infections. (2021).

Acknowledgements

See Supplementary section S1.

Author contributions

All authors have contributed to the publication, being variously involved in technology development, experimental protocol designs, experimental performance, data acquisition, statistical analysis, and manuscript writing.

Competing interests

The authors affiliated with Atomwise declare the existence of a financial competing interest.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-54655-z.

Correspondence and requests for materials should be addressed to

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

The Atomwise AIMS Program

Izhar Wallach[2], Denzil Bernard[2], Kong Nguyen[2], Gregory Ho[2], Adrian Morrison[2], Adrian Stecula[2], Andreana Rosnik[2], Ann Marie O'Sullivan[2], Aram Davtyan[2], Ben Samudio[2], Bill Thomas[2], Brad Worley[2], Brittany Butler[2], Christian Laggner[2], Desiree Thayer[2], Ehsan Moharreri[2], Greg Friedland[2], Ha Truong[2], Henry van den Bedem[2], Ho Leung Ng[2], Kate Stafford[2], Krishna Sarangapani[2], Kyle Giesler[2], Lien Ngo[2], Michael Mysinger[2], Mostafa Ahmed[2], Nicholas J. Anthis[2], Niel Henriksen[2], Pawel Gniewek[2], Sam Eckert[2], Saulo de Oliveira[2], Shabbir Suterwala[2], Srimukh Veccham Krishna PrasadPrasad[2], Stefani Shek[2], Stephanie Contreras[2], Stephanie Hare[2], Teresa Palazzo[2], Terrence E. O'Brien[2], Tessa Van Grack[2], Tiffany Williams[2], Ting-Rong Chern[2], Victor Kenyon[2], Andreia H. Lee[3], Andrew B. Cann[4], Bastiaan Bergman[5], Brandon M. Anderson[6], Bryan D. Cox[7], Jeffrey M. Warrington[8], Jon M. Sorenson[9], Joshua M. Goldenberg[10], Matthew A. Young[11], Nicholas DeHaan[12], Ryan P. Pemberton[13], Stefan Schroedl[14], Tigran M. Abramyan[11,15], Tushita Gupta[16], Venkatesh Mysore[17], Adam G. Presser[18], Adolfo A. Ferrando[19], Adriano D. Andricopulo[20], Agnidipta Ghosh[21], Aicha Gharbi Ayachi[22], Aisha Mushtaq[23], Ala M. Shaqra[24], Alan Kie Leong Toh[25], Alan V. Smrcka[26], Alberto Ciccia[27], Aldo Sena de Oliveira[28], Aleksandr Sverzhinsky[29], Alessandra Mara de Sousa[30], Alexander I. Agoulnik[31], Alexander Kushnir[32], Alexander N. Freiberg[33], Alexander V. Statsyuk[34], Alexandre R. Gingras[35], Alexei Degterev[36], Alexey Tomilov[37], Alice Vrielink[38], Alisa A. Garaeva[39], Amanda Bryant-Friedrich[40], Amedeo Caflisch[41], Amit K. Patel[35], Amith Vikram Rangarajan[42], An Matheeussen[43], Andrea Battistoni[44], Andrea Caporali[45], Andrea Chini[46], Andrea Ilari[47], Andrea Mattevi[48], Andrea Talbot Foote[49], Andrea Trabocchi[50], Andreas Stahl[51], Andrew B. Herr[52], Andrew Berti[40], Andrew Freywald[53], Andrew G. Reidenbach[54], Andrew Lam[55], Andrew R. Cuddihy[56], Andrew White[57], Angelo Taglialatela[19], Anil K. Ojha[58], Ann M. Cathcart[59], Anna A. L. Motyl[45], Anna Borowska[39], Anna D'Antuono[60], Anna K. H. Hirsch[61], Anna Maria Porcelli[62], Anna Minakova[48], Anna Montanaro[60], Anna Müller[41], Annarita Fiorillo[63], Anniina Virtanen[64], Anthony J. O'Donoghue[35], Antonio Del Rio Flores[51], Antonio E. Garmendia[65], Antonio Pineda-Lucena[66], Antonito T. Panganiban[67], Ariela Samantha[38], Arnab K. Chatterjee[68], Arthur L. Haas[69], Ashleigh S. Paparella[21], Ashley L. St. John[70], Ashutosh Prince[71], Assmaa ElSheikh[72], Athena Marie Apfel[57], Audrey Colomba[73], Austin O'Dea[74], Bakary N'tji Diallo[75], Beatriz Murta Rezende Moraes Ribeiro[76], Ben A. Bailey-Elkin[77], Benjamin L. Edelman[78], Benjamin Liou[52], Benjamin Perry[79], Benjamin Soon Kai Chua[80], Benjámin Kováts[81], Bernhard Englinger[59], Bijina Balakrishnan[82], Bin Gong[33], Bogos Agianian[21], Brandon Pressly[37], Brenda P. Medellin Salas[83], Brendan M. Duggan[35], Brian V. Geisbrecht[84], Brian W. Dymock[85], Brianna C. Morten[85], Bruce D. Hammock[37], Bruno Eduardo Fernandes Mota[76], Bryan C. Dickinson[86], Cameron Fraser[87], Camille Lempicki[88], Carl D. Novina[89], Carles Torner[90], Carlo Ballatore[35], Carlotta Bon[91], Carly J. Chapman[92], Carrie L. Partch[93], Catherine T. Chaton[94], Chang Huang[65], Chao-Yie Yang[95], Charlene M. Kahler[38], Charles Karan[27], Charles Keller[96], Chelsea L. Dieck[97], Chen Huimei[70], Chen Liu[98], Cheryl Peltier[77], Chinmay Kumar Mantri[70], Chinyere Maat Kemet[55], Christa E. Müller[99], Christian Weber[100], Christina M. Zeina[59], Christine S. Muli[101], Christophe Morisseau[37], Cigdem Alkan[33], Clara Reglero[19], Cody A. Loy[101], Cornelia M. Wilson[102], Courtney Myhr[31], Cristina Arrigoni[48], Cristina Paulino[39], César Santiago[103], Dahai Luo[22], Damon J. Tumes[104], Daniel A. Keedy[105], Daniel A. Lawrence[57], Daniel Chen[106], Danny Manor[71], Darci J. Trader[101], David A. Hildeman[52], David H. Drewry[107], David J. Dowling[108], David J. Hosfield[86], David M. Smith[109], David Moreira[110], David P. Siderovski[111], David Shum[112], David T. Krist[113], David W. H. Riches[78], Davide Maria Ferraris[114], Deborah H. Anderson[115], Deirdre R. Coombe[116], Derek S. Welsbie[35], Di Hu[71], Diana Ortiz[117], Dina Alramadhani[118], Dingqiang Zhang[119], Dipayan Chaudhuri[82], Dirk J. Slotboom[39], Donald R. Ronning[120], Donghan Lee[121], Dorian Dirksen[122], Douglas A. Shoue[123], Douglas William Zochodne[124], Durga Krishnamurthy[125], Dustin Duncan[126], Dylan M. Glubb[92], Edoardo Luigi Maria Gelardi[127], Edward C. Hsiao[128], Edward G. Lynn[129], Elany Barbosa Silva[130], Elena Aguilera[131], Elena Lenci[50], Elena Theres Abraham[132], Eleonora Lama[62], Eleonora Mameli[45], Elisa Leung[126], Emily M. Christensen[133], Emily R. Mason[134], Enrico Petretto[70], Ephraim F. Trakhtenberg[135], Eric J. Rubin[18], Erick Strauss[136], Erik W. Thompson[25], Erika Cione[137], Erika Mathes Lisabeth[138], Erkang Fan[139], Erna Geessien Kroon[76], Eunji Jo[112], Eva M. García-Cuesta[103], Evgenia Glukhov[35], Evripidis Gavathiotis[21], Fang Yu[140], Fei Xiang[141], Fenfei Leng[142], Feng Wang[143], Filippo Ingoglia[82], Focco van den Akker[71], Francesco Borriello[144], Franco J. Vizeacoumar[145], Frank Luh[146], Frederick S. Buckner[139], Frederick S. Vizeacoumar[53], Fredj Ben Bdira[147], Fredrik Svensson[73], G. Marcela Rodriguez[148], Gabriella Bognár[81], Gaia Lembo[149], Gang Zhang[150], Garrett Dempsey[51], Gary Eitzen[151], Gaétan Mayer[152], Geoffrey L. Greene[86], George A. Garcia[57], Gergely L. Lukacs[153], Gergely Prikler[81], Gian Carlo G. Parico[93], Gianni Colotti[47], Gilles De Keulenaer[154], Gino Cortopassi[37], Giovanni Roti[60], Giulia Girolimetti[62], Giuseppe Fiermonte[155], Giuseppe Gasparre[156], Giuseppe Leuzzi[19], Gopal Dahal[157], Gracjan Michlewski[158,159], Graeme L. Conn[160], Grant David Stuchbury[85], Gregory R. Bowman[161], Grzegorz Maria Popowicz[162], Guido Veit[153], Guilherme Eduardo de Souza[20], Gustav Akk[163], Guy Caljon[43], Guzmán Alvarez[164], Gwennan Rucinski[165], Gyeongeun Lee[112], Gökhan Cildir[166], Hai Li[27], Hairol E. Breton[167], Hamed Jafar-Nejad[168], Han Zhou[169], Hannah P. Moore[170], Hannah Tilford[165], Haynes Yuan[171], Heesung Shim[37], Heike Wulff[37], Heinrich Hoppe[75], Helena Chaytow[45], Heng-Keat Tam[172], Holly Van Remmen[173], Hongyang Xu[174], Hosana Maria Debonsi[175], Howard B. Lieberman[27], Hoyoung Jung[176], Hua-Ying Fan[177], Hui Feng[55], Hui Zhou[19], Hyeong Jun Kim[178], Iain R. Greig[179], Ileana Caliandro[180], Ileana Corvo[181], Imanol Arozarena[182], Imran N. Mungrue[183], Ingrid M. Verhamme[184], Insaf Ahmed Qureshi[185], Irina Lotsaris[186], Isin Cakir[57], J. Jefferson P. Perry[195], Jacek Kwiatkowski[85], Jacob Boorman[71], Jacob Ferreira[188], Jacob Fries[189], Jadel Müller Kratz[79], Jaden Miner[82], Jair L. Siqueira-Neto[35], James G. Granneman[190], James Ng[165], James Shorter[161], Jan Hendrik Voss[99], Jan M. Gebauer[132], Janelle Chuah[109], Jarrod J. Mousa[191], Jason T. Maynes[192], Jay D. Evans[193], Jeffrey Dickhout[194], Jeffrey P. MacKeigan[138], Jennifer N. Jossart[195], Jia Zhou[33], Jiabei Lin[161], Jiake Xu[196], Jianghai Wang[146], Jiaqi Zhu[197], Jiayu Liao[195], Jingyi Xu[195], Jinshi Zhao[198], Jiusheng Lin[199], Jiyoun Lee[200], Joana Reis[48], Joerg Stetefeld[77], John B. Bruning[201], John Burt Bruning[80], John G. Coles[202], John J. Tanner[167], John M. Pascal[29], Jonathan So[59], Jordan L. Pederick[80], Jose A. Costoya[110], Joseph B. Rayman[19], Joseph J. Maciag[52], Joshua Alexander Nasburg[37], Joshua J. Gruber[203], Joshua M. Finkelstein[55], Joshua Watkins[165], José Miguel Rodríguez-Frade[204], Juan Antonio Sanchez Arias[205], Juan José Lasarte[206], Julen Oyarzabal[205], Julian Milosavljevic[88], Julie Cools[154], Julien Lescar[22], Julijus Bogomolovas[35], Jun Wang[148], Jung-Min Kee[176], Jung-Min Kee[178], Junzhuo Liao[207], Jyothi C. Sistla[118], Jônatas Santos Abrahão[76], Kamakshi Sishtla[208], Karol R. Francisco[35], Kasper B. Hansen[209], Kathleen A. Molyneaux[71], Kathryn A. Cunningham[33], Katie R. Martin[138], Kavita Gadar[210], Kayode K. Ojo[139], Keith S. Wong[126], Kelly L. Wentworth[128], Kent Lai[82], Kevin A. Lobb[75], Kevin M. Hopkins[27], Keykavous Parang[211], Khaled Machaca[212], Kien Pham[98], Kim Ghilarducci[213], Kim S. Sugamori[126], Kirk James McManus[77], Kirsikka Musta[64], Kiterie M. E. Faller[45], Kiyo Nagamori[96], Konrad J. Mostert[136], Konstantin V. Korotkov[94], Koting Liu[214], Kristiana S. Smith[215], Kristopher Sarosiek[216], Kyle H. Rohde[217], Kyu Kwang Kim[218], Kyung Hyeon Lee[219], Lajos Pusztai[98], Lari Lehtiö[220], Larisa M. Haupt[25], Leah E. Cowen[126], Lee J. Byrne[102], Leila Su[146], Leon Wert-Lamas[89], Leonor Puchades-Carrasco[221], Lifeng Chen[86], Linda H. Malkas[187], Ling Zhuo[222], Lizbeth Hedstrom[223], Lizbeth Hedstrom[223], Loren D. Walensky[59], Lorenzo Antonelli[63], Luisa Iommarini[62], Luke Whitesell[126], Lía M. Randall[224], M. Dahmani Fathallah[225], Maira Harume Nagai[198], Mairi Louise Kilkenny[226], Manu Ben-Johny[19], Marc P. Lussier[213], Marc P. Windisch[112], Marco Lolicato[48], Marco Lucio Lolli[180], Margot Vleminckx[43], Maria Cristina Caroleo[227], Maria J. Macias[90], Marilia Valli[20], Marim M. Barghash[126], Mario Mellado[204], Mark A. Tye[228], Mark A. Wilson[199], Mark Hannink[229], Mark R. Ashton[85], Mark Vincent C.dela Cerna[121], Marta Giorgis[179], Martin K. Safo[118], Martin St. Maurice[230], Mary Ann McDowell[123], Marzia Pasquali[82], Masfique Mehedi[231], Mateus Sá Magalhães Serafim[76], Matthew B. Soellner[57], Matthew G. Alteen[232], Matthew M. Champion[123], Maxim Skorodinsky[233], Megan L. O'Mara[234], Mel Bedi[40], Menico Rizzi[114], Michael Levin[119], Michael Mowat[235], Michael R. Jackson[236], Mikell Paige[219], Minnatallah Al-Yozbaki[102], Miriam A. Giardini[130], Mirko M. Maksimainen[220], Monica De Luise[62], Muhammad Saddam Hussain[208], Myron Christodoulides[165], Natalia Stec[158], Natalia Zelinskaya[160], Natascha Van Pelt[43], Nathan M. Merrill[57], Nathanael Singh[105], Neeltje A. Kootstra[237], Neeraj Singh[238], Neha S. Gandhi[25], Nei-Li Chan[214], Nguyen Mai Trinh[22], Nicholas O. Schneider[230], Nick Matovic[85], Nicola Horstmann[239], Nicola Longo[82], Nikhil Bharambe[22], Nirvan Rouzbeh[209], Niusha Mahmoodi[21], Njabulo Joyfull Gumede[240], Noelle C. Anastasio[33], Noureddine Ben Khalaf[225], Obdulia Rabal[205], Olga Kandror[216], Olivier Escaffre[33], Olli Silvennoinen[64], Ozlem Tastan Bishop[75], Pablo Iglesias[110], Pablo Sobrado[241], Patrick Chuong[242], Patrick O'Connell[138], Pau Martin-Malpartida[90], Paul Mellor[53], Paul V. Fish[73], Paulo Otávio Lourenço Moreira[30], Pei Zhou[198], Pengda Liu[107], Pengda Liu[107], Pengpeng Wu[243], Percy Agogo-Mawuli[111], Peter L. Jones[244], Peter Ngoi[93], Peter Toogood[57], Philbert Ip[126], Philipp von Hundelshausen[100], Pil H. Lee[57], Rachael B. Rowswell-Turner[218], Rafael Balaña-Fouce[245], Rafael Eduardo Oliveira Rocha[76], Rafael V. C. Guido[20], Rafaela Salgado Ferreira[76], Rajendra K. Agrawal[58], Rajesh K. Harijan[21], Rajesh Ramachandran[246], Rajkumar Verma[247], Rakesh K. Singh[248], Rakesh Kumar Tiwari[249], Ralph Mazitschek[228], Rama K. Koppisetti[167], Remus T. Dame[147], Renée N. Douville[250], Richard C. Austin[194], Richard E. Taylor[123], Richard G. Moore[218], Richard H. Ebright[148], Richard M. Angell[73], Riqiang Yan[238], Rishabh Kejriwal[65], Robert A. Batey[126], Robert Blelloch[128], Robert J. Vandenberg[186], Robert J. Hickey[187], Robert J. Kelm Jr.[49], Robert J. Lake[177], Robert K. Bradley[251], Robert M. Blumenthal[106], Roberto Solano[46], Robin Matthias Gierse[252], Ronald E. Viola[157], Ronan R. McCarthy[210], Rosa Maria Reguera[245], Ruben Vazquez Uribe[253], Rubens Lima do Monte-Neto[30], Ruggiero Gorgoglione[155], Ryan T. Cullinane[223], Sachin Katyal[171], Sakib Hossain[105], Sameer Phadke[57], Samuel A. Shelburne[239], Sandra E. Geden[217], Sandra Johannsen[61], Sarah Wazir[220], Scott Legare[77], Scott M. Landfear[117], Senthil K. Radhakrishnan[118], Serena Ammendola[44], Sergei Dzhumaev[254], Seung-Yong Seo[141], Shan Li[143], Shan Zhou[168], Shaoyou Chu[134], Shefali Chauhan[255], Shinsaku Maruta[256,257], Shireen R. Ashkar[57], Show-Ling Shyng[117], Silvestro G. Conticello[149,257], Silvia Buroni[48], Silvia Garavaglia[114], Simon J. White[65], Siran Zhu[158,159], Sofiya Tsimbalyuk[258], Somaia Haque Chadni[142], Soo Young Byun[112], Soonju Park[112], Sophia Q. Xu[259], Sourav Banerjee[260], Stefan Zahler[222], Stefano Espinoza[91], Stefano Gustincich[91], Stefano Sainas[180], Stephanie L. Celano[138], Stephen J. Capuzzi[107], Stephen N. Waggoner[261], Steve Poirier[262], Steven H. Olson[236], Steven O. Marx[263], Steven R. Van Doren[167], Suryakala Sarilla[184], Susann M. Brady-Kalnay[71], Sydney Dallman[231], Syeda Maryam Azeem[105], Tadahisa Teramoto[264], Tamar Mehlman[105], Tarryn Swart[75], Tatjana Abaffy[265], Tatos Akopian[216], Teemu Haikarainen[64], Teresa Lozano Moreda[266], Tetsuro Ikegami[33], Thaiz Rodrigues Teixeira[175], Thilina D. Jayasinghe[120], Thomas H. Gillingwater[45], Thomas Kampourakis[267], Timothy I. Richardson[208], Timothy J. Herdendorf[84], Timothy J. Kotzé[136], Timothy R. O'Meara[268], Timothy W. Corson[208], Tobias Hermle[88], Tomisin Happy Ogunwa[256], Tong Lan[86], Tong Su[229], Toshihiro Banjo[269], Tracy A. O'Mara[92], Tristan Chou[42], Tsui-Fen Chou[143], Ulrich Baumann[132], Umesh R. Desai[118], Vaibhav P. Pai[119], Van Chi Thai[38], Vasudha Tandon[260], Versha Banerji[77], Victoria L. Robinson[65], Vignesh Gunasekharan[169], Vigneshwaran Namasivayam[99], Vincent F. M. Segers[43], Vincent Maranda[53], Vincenza Dolce[137], Vinícius Gonçalves Maltarollo[76], Viola Camilla Scoffone[48], Virgil A. Woods[105], Virginia Paola Ronchi[270], Vuong Van Hung Le[271], W. Brent Clayton[101], W. Todd Lowther[272], Walid A. Houry[126], Wei Li[273], Weiping Tang[207],

Wenjun Zhang[51], Wesley C. Van Voorhis[139], William A. Donaldson[230], William C. Hahn[59], William G. Kerr[274], William H. Gerwick[130], William J. Bradshaw[275], Wuen Ee Foong[276], Xavier Blanchet[277], Xiaoyang Wu[86], Xin Lu[123], Xin Qi[246], Xin Xu[84], Xinfang Yu[168], Xingping Qin[278], Xingyou Wang[223], Xinrui Yuan[95], Xu Zhang[279], Yan Jessie Zhang[83], Yanmei Hu[148], Yasser Ali Aldhamen[138], Yicheng Chen[71], Yihe Li[71], Ying Sun[52], Yini Zhu[123], Yogesh K. Gupta[280], Yolanda Pérez-Pertejo[245], Yong Li[168], Young Tang[65], Yuan He[40], Yuk-Ching Tse-Dinh[142], Yulia A. Sidorova[281], Yun Yen[146], Yunlong Li[282], Zachary J. Frangos[283], Zara Chung[22], Zhengchen Su[33], Zhenghe Wang[71], Zhiguo Zhang[27], Zhongle Liu[126], Zintis Inde[216], Zoraima Artía[164] & Abraham Heifets[2]

[2]Atomwise Inc., San Fransico, USA. [3]Amgen, Thousand Oaks, USA. [4]OpenAI, San Francisco, USA. [5]Model Medicines, La Jolla, USA. [6]Atomic.AI, San Francisco, USA. [7]Edifice Health, Inc., San Mateo, USA. [8]METiS Therapeutics, Cambridge, USA. [9]Genentech, San Mateo, USA. [10]US Navy Medical Service Corps Officer (2300/1810D), San Mateo, USA. [11]Totus Medicines, Inc., Emeryville, USA. [12]Cytokinetics, Inc., South San Francisco, USA. [13]Nurix Therapeutics, San Francisco, USA. [14]Amazon Alexa, Suite, USA. [15]The University of North Carolina at Chapel Hill Eshelman School of Pharmacy, Chapel Hill, USA. [16]Refibered Inc., Cupertino, USA. [17]NVIDIA, Santa Clara, USA. [18]Harvard TH Chan School of Public Health, Boston, USA. [19]Columbia University, New York, USA. [20]University of São Paulo, São Paulo, Brazil. [21]Albert Einstein College of Medicine, Bronx, USA. [22]Nanyang Technological University, Singapore, Singapore. [23]University of Washington, Seattle, USA. [24]Chan Medical School, University of Massachusetts, Worcester, USA. [25]Queensland University of Technology, Brisbane, USA. [26]University of Michigan Medical School, Ann Arbor, USA. [27]Columbia University Irving Medical Center, New York, USA. [28]Universidade Federal de Santa Catarina, Florianópolis, Brazil. [29]Université de Montréal, Montreal, Canada. [30]Instituto René Rachou-Fundação Oswaldo Cruz/Fiocruz Minas, Belo Horizonte, Brazil. [31]Herbert Wertheim College of Medicine, Biomolecular Science Institute, Florida International University, Miami, USA. [32]NYU Langone Health, New York, USA. [33]The University of Texas Medical Branch at Galveston, Galveston, USA. [34]University of Houston, Galveston, USA. [35]University of California, San Diego, USA. [36]School of Medicine, Tufts University, Medford, USA. [37]University of California, Davis, Davis, USA. [38]University of Western Australia, Crawley, Australia. [39]University of Groningen, Groningen, The Netherlands. [40]Wayne State University, Detroit, USA. [41]University of Zurich, Zürich, Switzerland. [42]Stanford University, Stanford, USA. [43]University of Antwerp, Antwerp, Belgium. [44]University of Rome Tor Vergata, Rome, Italy. [45]University of Edinburgh, Edinburgh, UK. [46]Department of Plant Molecular Genetics, Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CNB-CSIC), Madrid, Spain. [47]CNR (Italian National Research Council), Rome, Italy. [48]University of Pavia, Pavia, Italy. [49]University of Vermont, Burlington, USA. [50]University of Florence, Florence, Italy. [51]University of California, Berkeley, Berkeley, USA. [52]Cincinnati Children's Hospital Medical Center, Cincinnati, USA. [53]University of Saskatchewan, Saskatoon, Canada. [54]Broad Institute of MIT and Harvard, Cambridge, USA. [55]Boston University, Boston, USA. [56]CancerCare Manitoba Research Institute, Winnipeg, Canada. [57]University of Michigan, Ann Arbor, USA. [58]Wadsworth Center, New York State Department of Health and University at Albany, Albany, USA. [59]Dana-Farber Cancer Institute, Boston, USA. [60]University of Parma, Parma, Italy. [61]Helmholtz Institute for Pharmaceutical Research Saarland, Saarbrücken, Germany. [62]University of Bologna, Bologna, Italy. [63]Sapienza University of Rome, Rome, Italy. [64]Tampere University, Tampere, Finland. [65]University of Connecticut, Storrs, USA. [66]Centro de Investigación Médica Aplicada, Universidad de Navarra, Pamplona, Spain. [67]Tulane National Primate Research Center, Tulane University, Covington, USA. [68]Scripps Research, San Diego, USA. [69]Louisiana State University School of Medicine, New Orleans, USA. [70]Duke-NUS Medical School, Singapore, Singapore. [71]Case Western Reserve University, Cleveland, USA. [72]Oregon Health and Science University and Tanta University in Tanta, Tanta, Egypt. [73]University College London, London, UK. [74]Saint Louis University, St. Louis, USA. [75]Rhodes University, Makhanda, South Africa. [76]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil. [77]University of Manitoba, Winnipeg, Canada. [78]National Jewish Health, Denver, USA. [79]Drugs for Neglected Diseases Initiative (DNDi), Geneva, Switzerland. [80]The University of Adelaide, Adelaide, Australia. [81]Mcule, Budapest, Hungary. [82]University of Utah, Salt Lake City, USA. [83]The University of Texas at Austin, Austin, USA. [84]Kansas State University, Manhattan, USA. [85]UniQuest Pty Ltd, St Lucia, Australia. [86]University of Chicago, Chicago, USA. [87]Harvard University, Cambridge, USA. [88]University of Freiburg, Freiburg Im Breisgau, Germany. [89]Dana-Farber Cancer Institute and Harvard Medical School, Boston, USA. [90]IRB Barcelona, Barcelona, Spain. [91]Istituto Italiano Di Tecnologia, Genoa, Italy. [92]QIMR Berghofer Medical Research Institute, Herston, Australia. [93]University of California, Santa Cruz, Santa Cruz, USA. [94]University of Kentucky, Lexington, USA. [95]University of Tennessee Health Science Center, Memphis, USA. [96]Children's Cancer Therapy Development Institute, Beaverton, USA. [97]Columbia University Medical Center, New York, USA. [98]Yale School of Medicine, New Haven, USA. [99]University of Bonn, Bonn, Germany. [100]Ludwig-Maximilians-Universität München, Munich, Germany. [101]Purdue University, West Lafayette, USA. [102]Canterbury Christ Church University, Canterbury, UK. [103]National Centre for Biotechnology (CNB-CSIC), Madrid, Spain. [104]University of South Australia and SA Pathology, Adelaide, Australia. [105]CUNY Advanced Science Research Center, New York, USA. [106]The University of Toledo, Toledo, USA. [107]University of North Carolina at Chapel Hill, Chapel Hill, USA. [108]Boston Children's Hospital and Harvard Medical School, Boston, USA. [109]West Virginia University, Morgantown, USA. [110]Universidade de Santiago de Compostela, Santiago, Spain. [111]University of North Texas Health Science Center at Fort Worth, Fort Worth, USA. [112]Institut Pasteur Korea, Seongnam, South Korea. [113]Carle Illinois College of Medicine, Urbana, USA. [114]Università del Piemonte Orientale, Vercelli, Italy. [115]Saskatchewan Cancer Agency, Saskatoon, Canada. [116]Curtin University, Bentley, Australia. [117]Oregon Health and Science University, Portland, USA. [118]Virginia Commonwealth University, Richmond, USA. [119]Tufts University, Medford, USA. [120]University of Nebraska Medical Center, Omaha, USA. [121]University of

Louisville, Louisville, USA. [122]Dana Farber Cancer Institute, Boston, USA. [123]University of Notre Dame, Notre Dame, USA. [124]University of Alberta, Edmonton, Canada. [125]Cincinnati Childrens Hospital Medical Center, Cincinnati, USA. [126]University of Toronto, Toronto, Canada. [127]University of Piemonte Orientale, Vercelli, Italy. [128]University of California, San Francisco, San Francisco, USA. [129]St. Joseph's Healthcare Hamilton, and Hamilton Center for Kidney Research, McMaster University, Hamilton, Canada. [130]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, San Diego, USA. [131]Universidad de La República, Montevideo, Uruguay. [132]University of Cologne, Cologne, Germany. [133]Johnson University, Knoxville, USA. [134]Indiana University, Bloomington, USA. [135]School of Medicine, University of Connecticut, Farmington, USA. [136]Stellenbosch University, Stellenbosch, South Africa. [137]University of Calabria, Arcavacata, Italy. [138]Michigan State University, East Lansing, USA. [139]University of Washington, Washington, USA. [140]Weill Cornell Medicine-Qatar, Ar-Rayyan, Qatar. [141]Gachon University, Seongnam, South Korea. [142]Florida International University, Miami, USA. [143]California Institute of Technology, Pasadena, USA. [144]Boston Children's Hospital, Boston, USA. [145]Saskatchewan Cancer Agency and University of Saskatchewan, Saskatchewan, Canada. [146]Sino-American Cancer Foundation, Covina, USA. [147]Leiden University, Leiden, The Netherlands. [148]Rutgers University, Newark, USA. [149]Core Research Laboratory, ISPRO, Florence, Italy. [150]Caltech, Pasadena, USA. [151]University of Alberta, Edmonton, USA. [152]Montreal Heart Institute and Université de Montréal, Montreal, Canada. [153]McGill University, Montreal, Canada. [154]Antwerp University, Antwerp, Belgium. [155]University of Bari Aldo Moro, Bari, Italy. [156]Alma Mater Studiorum-University of Bologna, Bologna, Italy. [157]University of Toledo, Toledo, USA. [158]International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland. [159]Infection Medicine, University of Edinburgh The Chancellor's Building, Edinburgh, UK. [160]Emory University, Atlanta, USA. [161]University of Pennsylvania, Philadelphia, USA. [162]Helmholtz Zentrum München, Munich, Germany. [163]Washington University School of Medicine, St. Louis, USA. [164]CENUR Litoral Norte, Universidad de La República, Montevideo, Uruguay. [165]University of Southampton, Southampton, UK. [166]Centre for Cancer Biology, University of South Australia, Adelaide, Australia. [167]University of Missouri, Columbia, USA. [168]Baylor College of Medicine, Houston, USA. [169]Yale University, New Haven, USA. [170]Reno School of Medicine, University of Nevada, Reno, USA. [171]University of Manitoba and CancerCare Manitoba, Winnipeg, Canada. [172]Goethe University Frankfurt, Frankfurt, Germany. [173]Oklahoma Medical Research Foundation/Oklahoma City VA Medical Center, Oklahoma City, USA. [174]Oklahoma Medical Research Foundation, Oklahoma City, USA. [175]Department of Biomolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil. [176]Ulsan National Institute of Science and Technology, Ulsan, South Korea. [177]University of New Mexico Comprehensive Cancer Center, Albuquerque, USA. [178]Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea. [179]University of Aberdeen, Aberdeen, UK. [180]University of Turin, Turin, Italy. [181]Universidad de La República, CenUR LN, Montevideo, Uruguay. [182]Navarrabiomed-IdiSNA, Pamplona, Spain. [183]Independent, Los Angeles, USA. [184]Vanderbilt University Medical Center, Nashville, USA. [185]University of Hyderabad, Hyderabad, India. [186]University of Sydney, Sydney, Australia. [187]City of Hope Medical Center, Duarte, USA. [188]Weill Cornell Medicine, New York, NY 10065, USA. [189]University of Toledo College of Medicine and Life Sciences, Toledo, USA. [190]School of Medicine, Wayne State University, Detroit, USA. [191]University of Georgia, Athens, USA. [192]The Hospital for Sick Children, Toronto, Canada. [193]United States Department of Agriculture, Agricultural Research Service (USDA-ARS), Washington, DC, USA. [194]McMaster University, Hamilton, Canada. [195]University of California, Riverside, Riverside, USA. [196]The University of Western Australia, Perth, Australia. [197]The University of Connecticut, Storrs, USA. [198]Duke University School of Medicine, Durham, USA. [199]University of Nebraska-Lincoln, Lincoln, USA. [200]Sungshin University, Seoul, South Korea. [201]University of Adelaide, Adelaide, Australia. [202]University Toronto, Toronto, Canada. [203]University of Texas Southwestern Medical Center, Dallas, USA. [204]Centro Nacional de Biotecnologia/CSIC, Madrid, Spain. [205]Centro de Investigación Médica Aplicada, Pamplona, Spain. [206]Centro de Investigación Médica Aplicada, Universidad de Navarra, Pamplona, Spain. [207]University of Wisconsin-Madison, Madison, USA. [208]Indiana University School of Medicine, Indianapolis, USA. [209]University of Montana, Missoula, USA. [210]Brunel University London, London, UK. [211]Chapman University, Orange, USA. [212]Weill Cornell Medicine Qatar, Ar-Rayyan, Qatar. [213]Université du Québec À Montréal, Montreal, Canada. [214]National Taiwan University, Taipei, Taiwan. [215]Rhodes College, Memphis, USA. [216]Harvard School of Public Health, Boston, USA. [217]University of Central Florida, Orlando, USA. [218]University of Rochester, Rochester, USA. [219]George Mason University, Fairfax, USA. [220]University of Oulu, Oulu, Finland. [221]Instituto Investigación Sanitaria La Fe, Valencia, Spain. [222]Ludwig-Maximilians-University, Munich, Germany. [223]Brandeis University, Waltham, USA. [224]Universidad de La República, CENUR Litoral Norte, Montevideo, Uruguay. [225]Arabian Gulf University, Manama, Bahrain. [226]University of Cambridge, Cambridge, UK. [227]University of Magna Graecia, Catanzaro, Italy. [228]Massachusetts General Hospital, Boston, USA. [229]University of Missouri-Columbia, Columbia, USA. [230]Marquette University, Milwaukee, USA. [231]University of North Dakota, Grand Forks, USA. [232]Simon Fraser University, Burnaby, Canada. [233]CancerCare Manitoba Research Institute (CCMR), Winnipeg, Canada. [234]The University of Queensland, Brisbane, Australia. [235]University of Manitoba and CancerCare Manitoba Research Institute, Winnipeg, Canada. [236]Sanford Burnham Prebys, La Jolla, USA. [237]University of Amsterdam, Amsterdam, The Netherlands. [238]UConn Health, Farmington, USA. [239]The University of Texas MD Anderson Cancer Center, Houston, USA. [240]Walter Sisulu University, Mthatha, South Africa. [241]Virginia Tech, Blacksburg, USA. [242]University of Houston, Houston, USA. [243]Rutgers University, New Brunswick, USA. [244]University of Nevada, Reno, USA. [245]Universidad de León, León, Spain. [246]School of Medicine, Case Western Reserve University, Cleveland, USA. [247]School of Medicine, UConn Health, Farmington, USA. [248]University of Rochester Medical Center, Rochester, USA. [249]Chapman University School of Pharmacy, Irvine, USA. [250]University of Winnipeg/St. Boniface Research Centre, Winnipeg, Canada. [251]Fred Hutchinson Cancer Center, Seattle, USA. [252]Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Saarbrücken, Germany. [253]Technical University of Denmark, Kongens Lyngby, Denmark. [254]The City College of New York, New York, USA. [255]Children's Cancer, Therapy Development Institute (Cc-TDI), Beaverton, USA. [256]Soka University, Hachioji, Japan. [257]Institute of Clinical Physiology, National Research Council, Pisa, Italy. [258]Charles Sturt University, Bathurst, Australia. [259]Washington University, St Louis, USA. [260]University of Dundee, Dundee, UK. [261]Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, USA. [262]Montreal Heart Institute, Montreal, Canada. [263]Columbia University Vagelos College of Physicians and Surgeons, Columbia, USA. [264]Georgetown University, Washington, USA. [265]Duke University, Durham, USA. [266]Center for Applied Medical Research, University of Navarra, Pamplona, Spain. [267]King's College London, London, UK. [268]Precision Vaccines Program, Division of Infectious Diseases, Boston Children's Hospital, Boston, USA. [269]Fred Hutchinson Cancer Research Center, Seattle, USA. [270]Louisiana State University, Baton Rouge, USA. [271]Massey University, Palmerston North, New Zealand. [272]Wake Forest University School of Medicine, Winston-Salem, USA. [273]Central South University, Changsha, China. [274]SUNY Upstate Medical University, Syracuse, USA. [275]University of Oxford, Oxford, UK. [276]Goethe-University, Frankfurt, Frankfurt, Germany. [277]Institute for Cardiovascular Prevention (IPEK), Ludwig-Maximilians-Universität München, Munich, Germany. [278]Harvard T.H. Chan School of Public Health, Boston, USA. [279]School of Medicine, Boston University, Boston, USA. [280]University of Texas Health Science Center at San Antonio, San Antonio, USA. [281]University of Helsinki, Helsinki, Finland. [282]Wadsworth Center, NYSDOH, Albany, USA. [283]The University of Sydney, Sydney, Australia.

Annex D. Grants that have supported this research

I was awarded a FI doctoral fellowship from AGAUR (Agencia de Gestió d'Ajudes Universitaries i d'Investigació), reference 24A01100 for three years and after that, I had also a contract from la Fundació La Marató de TV3 and from the SGR 866-2021 grant and I also got granted a remote INSTRUCT internship INSTRUCT-ERIC (PID 1753).

The research was funded by the following projects:

Awarded to our lab:

- PID2021-122909NB-I00 from the Spanish Ministerio de Ciencia e Innovación MCIN/ AEI /https://doi.org/10.13039/ 501100011033/ and FEDER-A Way of making Europe.
- PDC_2021-121162-I00 from Spanish Ministerio de Ciencia e Innovación MCIN and by the European Union Next Generation EU/ PRTR.
- 201911 from Fundació La Marató de TV3.
- SGR 866-2021 from the AGAUR
- Producte (2023 PROD_00095) from the AGAUR.
- Access to Libraries of compounds were granted by EU-OPENSCREEN INSTRUCT-ERIC, (EU-OPENSCREEN-DRIVE and MedChem) funded by the European Union's Horizon 2020 Research and Innovation programme (grant agreement No 823893).
- MOSBRI-2022-97 from Molecular-Scale Biophysics Research Infrastructure (MOSBRI). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004806.
- Access to the EMBL Grenoble, is financed by INSTRUCT-ERIC (PID 1753).

Awarded to Aurora Martinez lab:

- UiB-node of NOR-OPENSCREEN and EU-OPENSCREEN
- The Foundation K.G. Jebsen and The Neuro-SysMed Center (RCN, Project No. 288164).

We also acknowledge the institutional support of the Department of Biomedicine, University of Bergen, the IRB Barcelona, the MCIN through the Centers of Excellence Severo Ochoa Award, the BBVA Foundation, the CERCA program of the Government of Catalonia, and the ICREA Program.

We also thank the PCB for support, the Protein Expression Core Facility, IRB Barcelona, for providing reagents, and the Automated Crystallography Platform, IRB Barcelona-CSIC, and the ALBA Synchrotron for help with the experiments performed at the BL13-

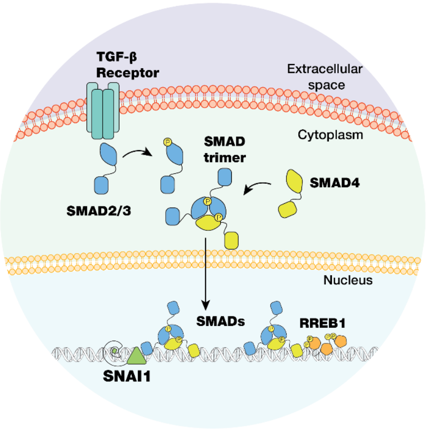
XALOC beamline. Access to the ALBA Synchrotron was through BAG proposal 2020094472. We also thank the ESRF for their help during our sessions at the Bio-SAXS BM29 synchrotron beamline and MOSBRI partners in the ProLinC facility in the University of Linköping (Dr. Dean Derbyshire) and in the SPC facility of EMBL Hamburg (Dr. Maria Marta Garcia Alai, Dr. Stephan Niebling and Dr. Angelica Struve) for access and assistance to state-of-the-art equipment to perform single-molecule biophysics assays. Special thanks go to our collaborators Dr. José A. Márquez and Dr. Anne-Sophie Humm at the High Throughput Facility Lab (HTX), EMBL Grenoble, for their help in setting up the crystallization for SMAD4 and complexes and also with the data processing and analysis.

Annex E. Thesis summary

Introduction

For as long as records have been kept, from ancient communities to the present, people have sought to relieve pain, prevent and treat infections, and alleviate and cure the symptoms of disease. Nowadays, therapeutic options have evolved to the point where we have numerous tools to treat a wide range of diseases. These tools include drugs of synthetic and natural origin, antibodies, and other biologics, as well as novel therapies, including the use of CRISPR technology and gene-editing tools to modify genes with harmful mutations, CAR-T cell therapy, new vaccines, and many others that are revolutionizing the field of medicine every day.

There are still unmet medical needs that require the identification of new -or complementary- treatments at a cost that is not prohibitive for the public health system. The identification of new compounds with pharmacological applications typically requires time and a substantial investment. However, if successful, large-scale production of the compounds could be less expensive than other alternatives, offsetting the initial economic investment, and facilitating its commercial production and use worldwide at an affordable cost. These needs include finding new treatments for cancer patients who have developed resistance to approved drugs and are running out of pharmacological options, or for individuals suffering from rare diseases, to name just a few. Rare diseases, in fact, are often overlooked by pharmaceutical companies because of the small number of people affected, the limited knowledge of the disease, and the lack of correlation between the observed phenotypes and the molecular basis.



Summary figure 1. SMAD signaling and interactions: Snail1 and RREB1.

With this in mind, we set out to explore the possibility of identifying molecules that could modulate the TGF-β signaling pathway, one of the seven signaling pathways conserved across Metazoan, combining the expertise of our lab at the IRB Barcelona led by the ICREA research Prof. Maria J. Macias with this biological system, and that of Prof. Aurora Martinez, at the University of Bergen, related to the identification of small compounds with pharmacological activity applying complementary biophysical techniques. TGF-β signaling in brief, includes a family of cytokines and membrane receptors that respond to these cytokines and, in the canonical pathway, a family of transcription factor proteins that act as the messengers of the receptor signals in the nucleus (**Summary figure 1**). This family of proteins are known as SMAD (Suppressor of Mothers against Decapentaplegic) proteins. SMAD-driven signaling is involved in many essential aspects of metazoans life, including embryo development, cell homeostasis, tumor suppressor, etc. Because of its importance to the proper functioning of our cells, this signaling network is tightly regulated. Unfortunately, this signaling network is not error-free, and mutations in SMAD proteins, particularly within SMAD4, have been associated with human diseases such as cancer and rare diseases **(Massagué and Sheppard, 2023)**.

SMADs are composed of an N-terminal domain that interacts with DNA, a linker, and a C-terminal domain that participates in protein-protein interactions **(Macias, Martin-Malpartida and Massagué, 2015)**. Both of these domains are unique to SMAD proteins. Another characteristic of SMAD proteins is to associate among them to form heterotrimers, which is the core transcriptional unit. The functional capabilities of the core SMAD complex are further modulated by the formation of SMAD complexes with other proteins (co-activators and repressors, ubiquitin ligases, kinases, phosphatases, and chromatin remodelers, to name a few) that fine-tune the functional properties of the SMAD-driven signaling system according to cellular needs **(Guca *et al.*, 2018; Aragón *et al.*, 2019; Su *et al.*, 2020)**.

While major therapeutic strategies to tackle TGF-β pathway are being focused to modulate the membrane receptor function or to inhibit the hormone activation **(Akhurst, 2017; Cho *et al.*, 2020; Liu, Ren and Ten Dijke, 2021; Yap *et al.*, 2021; Shi *et al.*, 2022)**, no therapeutic strategies have been tested in preclinical or clinical assays targeting SMAD proteins. Targeting SMAD4 can be of special interest since it is the most mutated element in the SMAD driven TGF-β pathway in primary tumors, specially in pancreatic and gastrointestinal tract cancers, and has key roles in advanced cancer stages, fibrosis and rare diseases. Patients such as those with Juvenile Polyposis Syndrome (JPS) or Hereditary Hemorrhagic Telangiectasia (HHT) **(Miyaki and Kuroki, 2003; Cao, Plazzer and Macrae, 2023)** usually have alterations in the proper function of epithelial tissue in various organs. The SMAD4 variants associated with these epithelial disorders, which accumulate mainly in the MH2 domain of the protein, cause inhibition of SMAD complex formation. Individuals with Myhre syndrome (MyS) have specific SMAD4 point mutations associated with stabilization of SMAD proteins. Remarkably, variants linked to rare diseases are often found as well in cancer patients.

In addition to these applied aims, we also planned to contribute to a better understanding of the molecular mechanisms of Epithelial to Mesenchymal transition (EMT), a phenotypic characteristic required during development and tissue repair but that can promote cancer invasion and metastasis in scenarios associated with disease. EMTs are driven by specialized signaling events that activate the expression of a set of transcription factors (EMT TFs) that repress epithelial genes and induce the expression of mesenchymal features **(Batlle *et al.*, 2000; Cano *et al.*, 2000)**. For our studies, we have selected a specialized effector of RAS/MAPK signaling, RREB1 (RAS response element binding protein 1) that also receives inputs from TGF-β to induce EMT and metastatic outgrowth in carcinoma cells **(Janda *et al.*, 2002; David *et al.*, 2016; Deng *et al.*, 2020; Su *et al.*, 2020)**. RREB1 is a large multi-Zinc finger (abbreviated as ZF) protein, four times longer than average protein sequences in eukaryotes (book.bionumbers.org). The ZF domains are the most abundant DNA binding structures found in eukaryotic transcription factors, present in more than 800 proteins in the human proteome **(Wolfe, Nekludova and Pabo, 2000; Najafabadi *et al.*, 2015)**. The ZFs of RREB1 are grouped into three main clusters, separated by large intervening regions lacking other known structured domains. Our contribution has been to analyse the interactions between the cluster of ZFs located at the C-terminal part of the protein and specific DNA motifs. This project is carried out as a collaboration with the laboratory of

Dr. Joan Massagué (Cancer Biology and Genetics Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA).

Working hypotheses of this work

Many SMAD mutations in cancer are thought to correlate with a loss of function role of SMAD4 proteins **(Miyaki and Kuroki, 2003; Chacko *et al.*, 2004; Massagué and Sheppard, 2023)**, whereas in Mhyre syndrome (MyS), a rare disease affecting embryo development and multiple organs, these mutations are correlated with a gain of function role which lead to increased SMAD4 protein levels and decreased ubiquitination of the protein in patient cell lines **(Le Goff *et al.*, 2011; Caputo *et al.*, 2014)**. Thus, as SMAD proteins form quaternary structures, we have hypothesized that some of these MyS mutations and cancer variants might affect the stoichiometry of the SMAD complexes, giving rise to transcriptional complexes of modified selectivity and affinity for DNAs and cofactors. We also hypothesized that these mutations might affect the stability of these SMAD heterocomplexes, modifying the duration of the transcription activation of specific genes and giving rise to diseases.

Driven by the urgent societal need to find new treatments for cancer patients and also for individuals with Myhre syndrome and other rare diseases, we proposed the transcription factor SMAD4 as a target for drug discovery. If we could find small molecules that interact with SMAD4, these molecules could be developed either as research tools or as molecules with potential pharmaceutical application, depending on their specific action. We also planned to test drugs already on the market for drug repurposing, an option that will allow us a faster path to the clinic if effective compounds are found, avoiding the need for time- and cost-demanding toxicity studies.

The results section contains three chapters. The first one is focused on studying the quaternary structures of SMAD proteins and how a few selected disease-associated mutations in SMAD4 affect the tertiary and quaternary structure of SMAD complexes. The second chapter has been dedicated to identify compounds that interact with SMAD4 and in studying how they modulate SMAD interactions. The third chapter includes our studies of the C-terminal region of RREB1 and its DNA binding function.

General objectives: The ultimate goal of this work is to obtain new knowledge and illustrate key steps in the transforming growth factor β (TGF-β) pathway using molecular, biophysical, structural and chemical biology techniques. The innovative goal is to advance in the development of therapies for cancer and rare diseases, by identifying vulnerable sites in the involved proteins, notably SMAD4, and identifying modifying compounds.

Specific objectives: Studying the quaternary structures of SMAD proteins and how these structures are affected by a few selected disease associated-variants in SMAD4.Using high throughput screening (HTS) of libraries of compounds, identify molecules that interact with SMAD4, including a set of FDA/EMA approved drugs in the context of a drug repurposing screening campaign for cancer, fibrosis and rare diseases. To elucidate the specific contacts with DNA to reveal the binding preferences of the RREB1 C-terminal ZFs, using a combination of binding assays and atomic resolution techniques. With our objectives in mind, we designed a set of recombinant protein constructs that were expressed, purified and studied through different single-molecule biophysical assays and characterized through structural biology approaches. In this study, we focused specially into SMAD MH2 domain constructs, which are more stable and reproducible than full-length proteins, and in several variants associated with diseases, as detailed below. In addition, we have also expressed and purified the C-terminal region of the RREB1 protein for DNA and protein binding and for structural analyses.

To get accounted on specific techniques that were either new to me or to our laboratories, I participated in short visits to facilities in Europe founded by competitive applications. Our laboratory also got granted an EU-OPENSCREEN Drive project, and thanks to the visits I made to the laboratory of Prof. A. Martinez (University of Bergen in Norway) I learned how to perform and analyze HTS of large libraries of compounds in a systematic and reproducible manner. Thanks to INSTRUCT ERIC and MOSBRI initiatives, I was able to learn high throughput X-ray Crystallography at the HTX lab of EMBL Grenoble with Dr. J. Marquez (France), and protein-protein interactions (PPIs) characterization trough single-molecule biophysics at the Sample Preparation and Characterization Facility (SPC) of the EMBL Hamburg (Germany) and the Protein folding and Ligand Interaction Core facility (ProLinC) of the University of Linköping (Sweden).

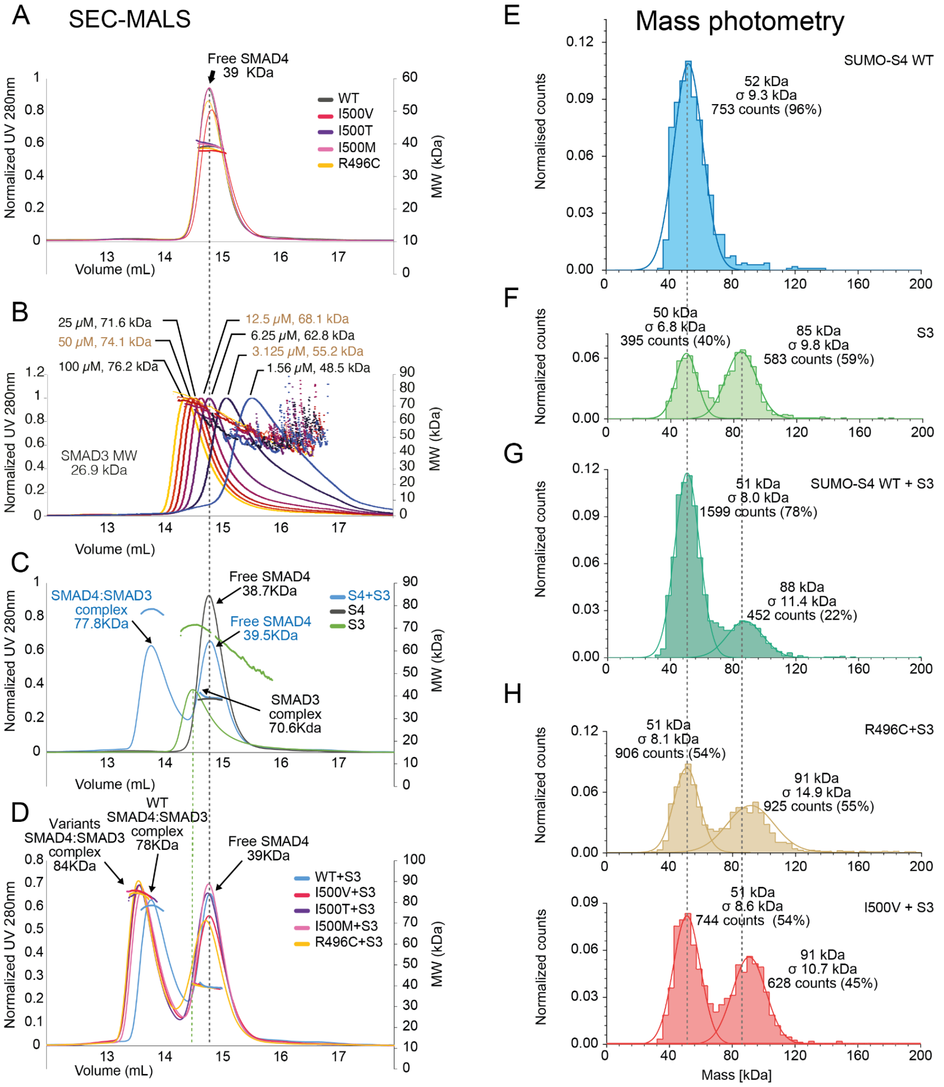
For this analysis we have expressed and purified the MH2 domains of SMAD4, SMAD2/3 and SMAD1/5 as well as mutations of the MH2 domain associated with diseases. The latter include all mutations described for the MyS, two more rare diseases (JPS and HHT) and also mutations identified in cancer patients.

Variants of the MyS and the set-up of the biophysical approaches

SMAD4 variants have been described in individuals with Myhre syndrome in 2011. They also found that individuals with these mutations had decreased SMAD4 ubiquitination and increased levels of both SMAD4 and activated R-SMADs proteins. We have expressed and purified each of these variants (R496C and I500V/M/T) and analyzed the effects of these mutations in the protein and complex formation and stability using different biophysical techniques, in solution. We have observed that each of the four mutations increases the affinity of interaction between the SMAD4 and R-SMADs with respect to the WT protein, which may also be translated into a functional gain of the signaling system in cells.

To characterize the effect of the mutations, we have applied two different single-molecule biophysical strategies: Size Exclusion Chromatography with Multi-Angle Light Scattering and Mass Photometry (abbreviated as SEC-MALS and MP, respectively). Through these assays, we observed an increase in the amount of complex formation of SMAD4 MyS variants with R-SMADs when compared with the wild-type protein. In these experimental conditions, SMAD complexes behave as a single pick with multiple species in equilibrium including monomers, dimers and trimers. SMAD4 behaves as a monomeric protein in solution by its own and when forming complexes with R-SMADs such oligomers do not include more than one SMAD4 unit.

In SEC-MALS, WT and MyS complexes with SMAD3 differ in elution volume and average calculated molecular mass (**Summary figure 2**). This phenomenon is produced by a different ratio of the SMAD complexes in the selected experimental conditions. MyS variants, when mixed with SMAD3, present an equilibrium more displaced towards heterotrimeric complex, compared to the WT protein (**Summary figure 2D**). In MP studies, we could accurately measure the number of SMAD complexes in solution, which allowed us to see an increment of heterotrimeric complexes ($MW_{theorical}$=92 kDa) when using MyS variants (**Summary figure 2E-H**). Through isothermal titration calorimetry (ITC) we explained how these changes were caused by variations in the thermodynamic parameters of binding and affinity, which differ in 3- and 5- fold compared with the WT protein.



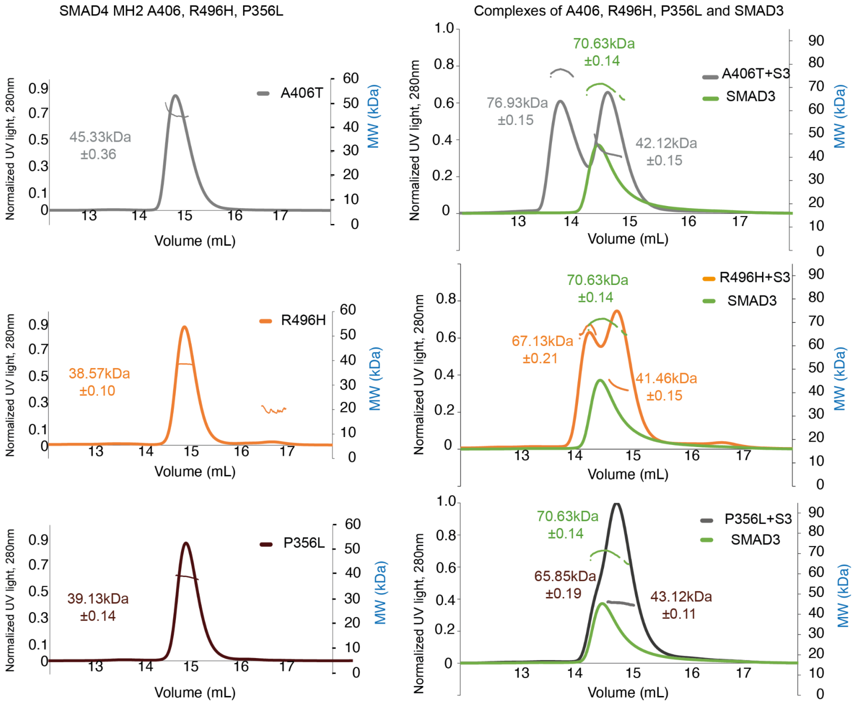
Summary figure 2. Complexes of WT and MyS SMAD4 variants with SMAD3. A. Size-exclusion chromatography coupled to multi-angle light scattering detector (SEC-MALS) data for SMAD4 WT and MyS variants show a single peak of a MW that corresponds to a monomer. B. Data corresponding to SMAD3 indicates the presence of dimers and trimers and the absence of monomers even at the lowest concentration. C. The complex between SMAD4 and SMAD3 shows the heterotrimer complex as well as a peak corresponding to the excess of unbound SMAD4. D. Compared to the WT protein, the MyS variants elute as a peak with a higher average molecular weight, indicating that the complex equilibrium is shifted toward the hetero-trimer formation. Mass photometry (MP) was used to quantify the number of particles corresponding to heterotrimeric complexes (right). E-F. Controls of the SMAD4 and SMAD3 proteins in the free state. G. A mixture of SMAD4 and SMAD3 reveals the presence of heterotrimers. H. Complexes with MyS variants show two times more heterotrimer particles compared to the WT scenario (G). Final protein concentrations in MP assays were 5.26 nM (SMAD3) and 2.63 nM (SMAD4). The monomeric constructs used in SEC-MALS and MP experiments had a molecular weight of 38.25 kDa for SUMO-SMAD4 314-552 and 26.88 kDa for SMAD3 189-425 DVD. The SMAD4:SMAD3:SMAD3 heterotrimer weighs 92 kDa, while the SMAD3 homotrimer weighs 80.6 kDa.

Using nanoDSF, we also observed that, upon SMAD4 binding, R-SMADs are stabilized. When comparing WT to MyS variants, we measured an increase of temperature of stabilization in the presence of these disease variants for all TGF-β/Nodal and BMP activated R-SMADs.

In SEC-MALS and MP techniques, we worked in conditions that stimulate the dissociation of SMAD complexes. As we were aware of this limitation, we set out to investigate if the differences in complex dissociation rate could be quantified through Surface Plasmon Resonance (SPR). The analysis allowed us to measure the effect of SMAD4 mutations on SMAD2/SMAD3 binding, using the latter as analytes. Using this approach, we observed a 5 to 6-fold increase in complex half-life time of mutated complexes compared to the WT.

Variants identified in cancer patients and in rare diseases, such as JPS and HHT

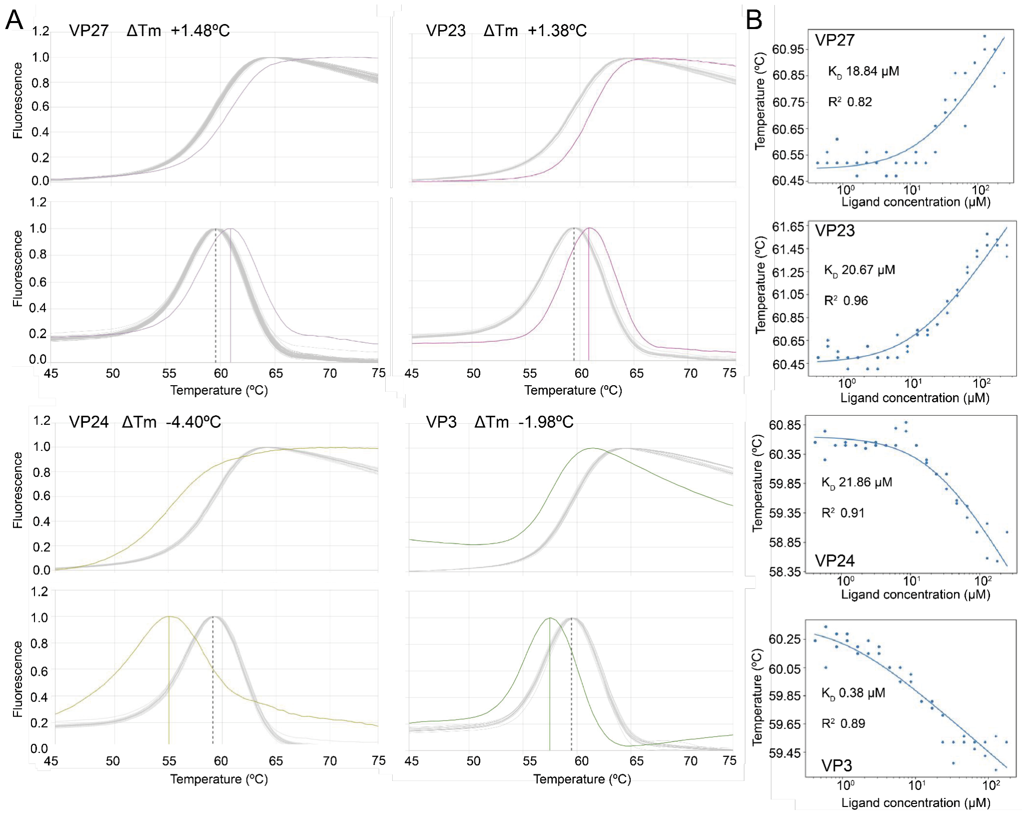
In order to characterize SMAD4 MH2 domain variants, we selected different residues of the MH2 domain which are mutated in patient samples, both in cancer patients and in individuals affected with rare diseases, such as JPS and HHT **(Cao, Plazzer and Macrae, 2023)**. Using the same biophysical approach as for the analysis of MyS variants (SEC-MALS), we aimed to find differences between the capability of each mutation to modulate SMAD complex formation. As a conclusion of our experiments, we observed how selected SMAD4 variants were able to disrupt in different degrees its binding with SMAD3. Based on the chromatographic profile of SEC-MALS data, we proposed a classification of the studied variants in three groups (**Summary figure 3**). Group 1 (G1), clearly form heterotrimers in presence of SMAD3 and behave similarly as the WT protein (including A406T, K428T and R515T variants). Group 2 (G2) has an evident loss of affinity that results in an elution of a pick of lower molecular weight compared to the G1 or WT complexes (including G386D and R496H variants). Finally, group 3 (G3) does not produce visible or detectable SMAD4-SMAD3 complexes in the applied conditions (R361G, P356L and D351G variants).



Summary figure 3. SMAD4 complex formation analyzed using SEC-MALS. (Left) Profiles corresponding to three variants, A406T, R496H and P356L. (Right) Complexes of these variants and SMAD3. SMAD4 samples were measured at 50 µM and SMAD3 at 25 µM.

In addition, we performed a structural study of the R496H variant, classified in G2 and found in the same residue as MyS variant R496C. R496H produces a partial loss of secondary structure in alpha-helix 4 (H4) of SMAD4 MH2 domain (**Figure 3B**). This produces a change in the orientation of the side chains in key residues for R-SMAD binding to SMAD4, like D493. This type of structural understanding on single point mutations in proteins is key to observe mutational effects, but also to contribute into protein structure databases that will be used to further develop *in-silico* tertiary and quaternary structure predictions. We wish to remark that at present, advanced programs as Colabfold, that predict tertiary structures of proteins, lack the accuracy to predict the effect of these variations at an atomic resolution **(Terwilliger *et al.*, 2024)**.The EU-OPENSCREEN drive project gave us access to a large library of compounds, specifically their Pilot and Diversity Libraries. This selection of compounds includes different sets of ligands, from FDA/EMA approved compounds to novel scaffolds with no reported pharmacological application to date, allowing a representative screening of the market chemical space. To identify binders of the MH2 domain of SMAD4, we screened a total of 100,037 compounds using Differential Scanning Fluorimetry (DSF), a method that monitors changes in the melting temperature (T_m) as a metric for binding compounds. The rationale is that binding of a small molecule could stabilize or destabilize the protein, resulting in a change in melting temperature. Compounds that produced a change in T_m greater than one to five times the standard deviation were considered hits. These hits were further validated by dose-response assays (DRA). To facilitate the identification of binders based on changes in T_m, we developed HTSDSF Explorer, a software program (fully open to the public) that facilitates the analysis and presentation of results. It also implements a tool to easily study DSF dose-response experiments, which is able to calculate the observed dissociation constant (K_D) **(Martin-Malpartida *et al.*, 2022)**.

From this analysis, we identified 462 hits (0.47% hit rate), from which 185 were validated in DRA (**Summary figure 4A,B**). We further validated the best hits with an orthogonal assay using spectral shift and Temperature-Related Intensity Change (TRIC).



Summary figure 4. Protein-ligand binding experiments between SMAD4 constructs and selected Tm modulators. A. Unfolding profile (TOP) and first derivative (BOTTOM) of SMAD4 272-552 incubated with 250µM of VP27, VP23, VP24 or VP3 compound. References, shown in gray, have a low standard deviation which allows the selection of these ligands as hits. B. DSF dose-response curve and K_D fitting using HTSDSF Explorer. Stabilizer (VP27 and VP23) and destabilizers (VP24 and VP3) are shown.

We also screened the Prestwick Chemical Library, for drug repurposing. In this screen we used up to three different SMAD4 variants (WT, R361G, R496C and I500V) with the idea to find variant specific compounds. We identified several hits for WT (12, 0.93% hit rate), I500V (21, 2.2% hit rate), R496C (30, 3.1% hit rate), and R361G (10, 1% hit rate) SMAD4 variants. Dose-response assays with these hits using DSF resulted in a final selection of 13 candidates. Nearly all the validated hits were able to interact to some extent with all protein constructs tested, except one that interacts almost exclusively with the R361G variant over the range of affinities we evaluated.

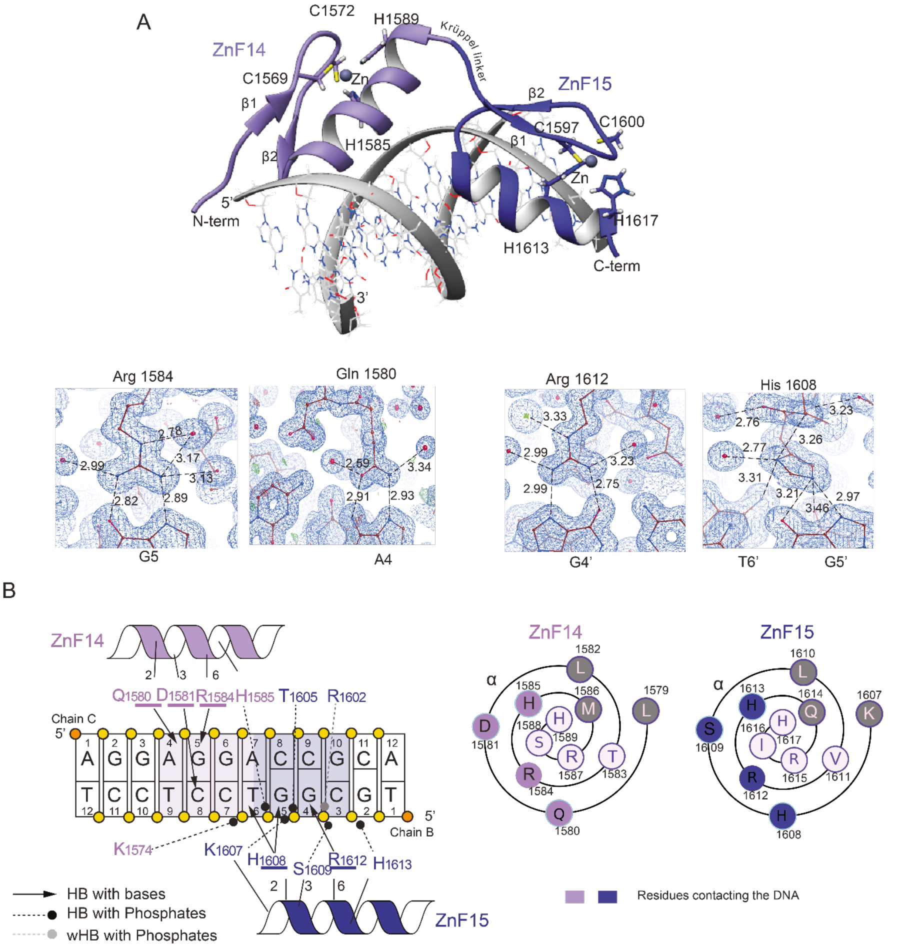
To advance the hit-to-lead process, we are attempting to obtain the protein-ligand complex structure of a small set of validated hits using X-ray crystallography. We have succeeded in determining the interaction site for one of the ligands, and we are currently optimizing the crystallization conditions to obtain higher resolution datasets to be certain of other ligand-protein contacts.

DNA binding recognition of the C-terminal ZF cluster of the RREB1 protein

This cluster is affected by splicing events. Two ZF 12-12' and 14-15 pairs are present in the isoform ***α*** and are highly conserved in *Drosophila,* whereas ZF 12' is absent in isoform **β**, the most commonly used isoform in the literature. ZF14-15 is also present in all isoforms except in the ***δ***. The isoform **ζ** only contains ZF 12', 13 and the 14-15 pair. Of all these pairs, the 14-15 one is almost 100% conserved in vertebrates **(Deng *et al.*, 2020)** and 53% identical to *Drosophila* hindsight **(Zhang, Zhao and Edenberg, 1999; Pickup, Ming and Lipshitz, 2009)**. Prior to the structural studies, we explored the binding preferences of the ZF pairs. A recombinant construct spanning ZFs 14-15 was prepared using standard protocols, yielding a highly soluble protein. However, constructs containing the ZF12 were prone to aggregation/precipitation, perhaps due to the presence of additional Cys and His residues, and due to this limitation, we focused the studies using the ZF 14-15 pair. For the DNA binding assays, we used a native DNA sequence derived from the SerpinE1 promoter containing the GGTCCT motif, which corresponds to a region bound in ChIP-seq experiments **(Su *et al.*, 2020).** Using isothermal titration calorimetry, we quantified the affinity of the recombinant 14-15 ZF pair, being in the nanomolar range (69.3 ± 26.3nM).

The best diffracting crystals were obtained with the 14-15 pair and a 12-mer dsDNA containing the GGTCC motif in the middle of the sequence. The structure of the complex has been refined at 1.15 Å resolution. The crystallographic asymmetric unit contains a copy of a protein-DNA complex in which each ZF makes specific interactions with half of the DNA motif and the pair wraps around almost all the DNA.

ZF14 binds to the second part of the 4-GGTCC-8 site, through specific base contacts between Gln1580 and Arg1584 (in the α-helix) and Guanine 5 and Adenine 4 nucleotides in the complementary strand, and from Asp1581 with Cytosine 8, in the primary strand. Both His1585, which also coordinates Zn, and Lys1574 (located at the second β-strand) contact the backbone DNA. ZF15 interacts with the first part of the 4-GGTCC-8 site. In this case, there are specific contacts between Arg1612 and Guanine 4 and His1608 with Guanine 5 and Thymine 6 bases (the latter with a suboptimal geometry) in the primary strand. In addition, due to a bend of the DNA, the protein can make abundant contacts with the backbone (phosphate groups), including interactions from Arg1602 residue located in the second β-strand as well as from and Thr1605 and from Ser1609 and His1613 residues in the α-helix itself. These contacts are indicated in **Summary figure 5A,B**. Overall, the observed pattern of contacts confirms that the 14-15 pair selects DNA motifs containing the AGx[A/T]CC sequences, consistent with motifs previously identified in the literature. This structure provides the first atomic description of how RREB1 recognizes specific DNA motifs. We are currently working to determine the role of other ZFs present in DNA binding, as this protein is found to interact with several loci genome wide, and we are also preparing a manuscript describing these results.



Summary figure 5. Complex structure of the ZF 14-15 pair bound to the GGTCCT-motif. A. A diagram of the protein-DNA complex together with electron density maps for the key contacts of both ZFs with the DNA. B. I Contacts with bases and backbone DNA.

The experimental work collected in this thesis provides a biophysical, structural- and chemical biology perspective of SMAD complexes, and has advanced the process of drug discovery targeting SMAD4 to find a pharmacological solution to diseases and syndromes that so far do not have an efficient treatment. We have also begun to elucidate the structural basis of DNA recognition of the RREB1 protein, a SMAD cofactor that promotes EMT processes and drives metastatic programs.

Conclusions

The specific conclusions can be summarized as follows:

1. Regarding the first objective,

    a. We have established a protocol to analyze how SMAD4 variants interact with R-SMADs.

    b. SMAD4 Myhre Syndrome variants form more stable complexes with R-SMADs than the WT, whereas cancer variants display different profiles depending on the specific mutation.

    c. We have determined the structure of the R496H variant, which reveals the effects of the mutation in the fold. These effects help understand how this point mutation affects the association with R-SMADs. These effects cannot be predicted using available software, strengthening the importance of having experimental data to establish structure-function relationships.

2. Regarding the second objective,

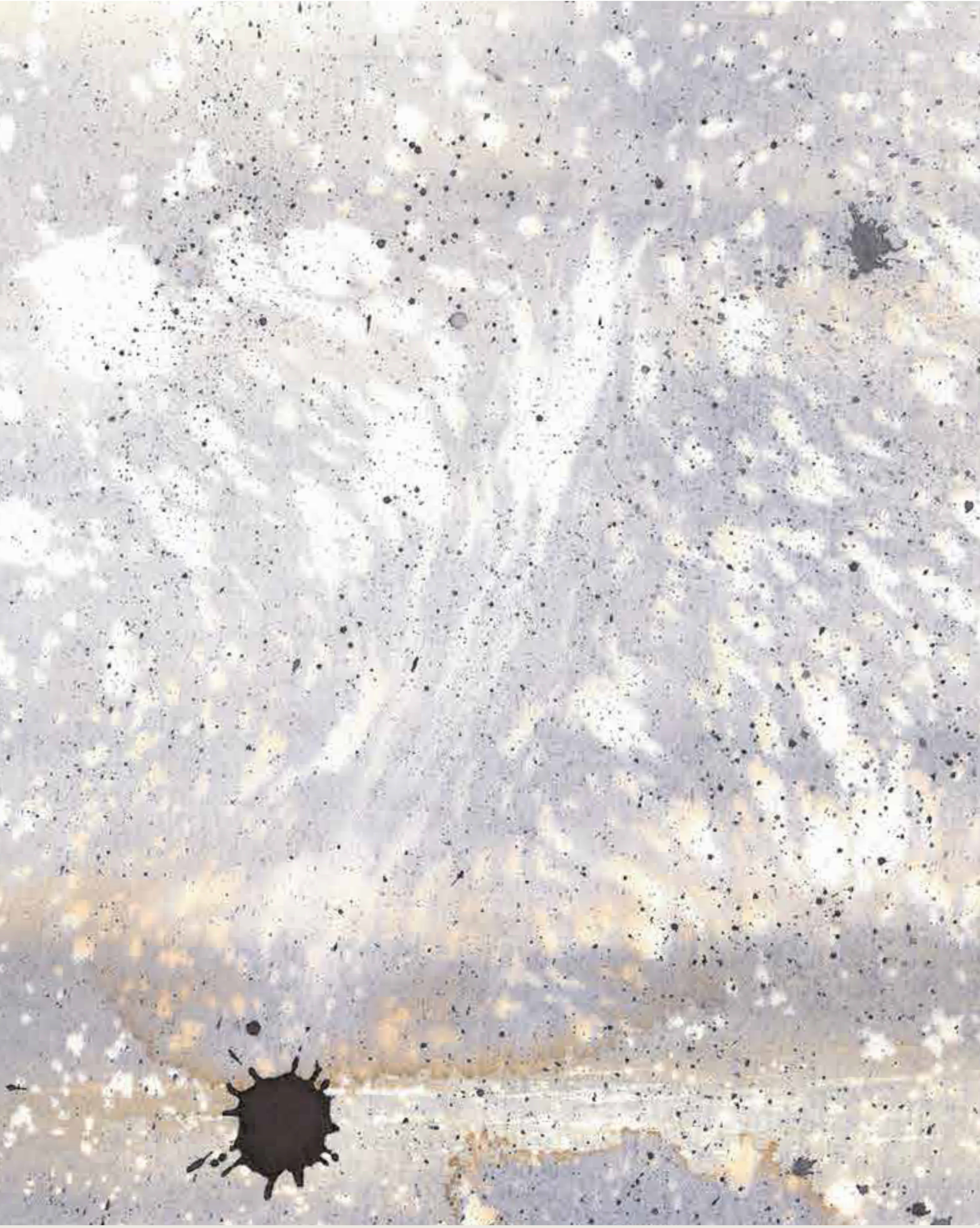
    a. Our HTS campaign using DSF has provided the first hit binders for SMAD4.

    b. We have also identified FDA/EMA approved compounds that have been validated as hits that may have a rapid path to the clinic for the benefit of patients suffering from cancer, fibrosis and/or rare diseases.

3. Regarding the third objective,

    a. RREB1 is a multi ZF transcription factor involved in EMT processes. We have found that the ZF 14-15 pair binds the GGTCCT motif with strong affinity.

    b. We have also elucidated the key residues in the protein and the specific nucleotides that participate in the recognition.



INSTITUT
DE RECERCA
BIOMÈDICA

IRB
BARCELONA

UNIVERSITAS BERGENSIS

UNIVERSITAT DE BARCELONA