



# LUND UNIVERSITY

## Structure and energetics of molecular recognition in galectin-3-ligand interactions

Kumar, Rohit

2019

[Link to publication](#)

*Citation for published version (APA):*

Kumar, R. (2019). *Structure and energetics of molecular recognition in galectin-3-ligand interactions*. Lund University, Faculty of Science.

*Total number of authors:*

1

### General rights

Unless other specific re-use rights are stated the following general rights apply:

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: <https://creativecommons.org/licenses/>

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

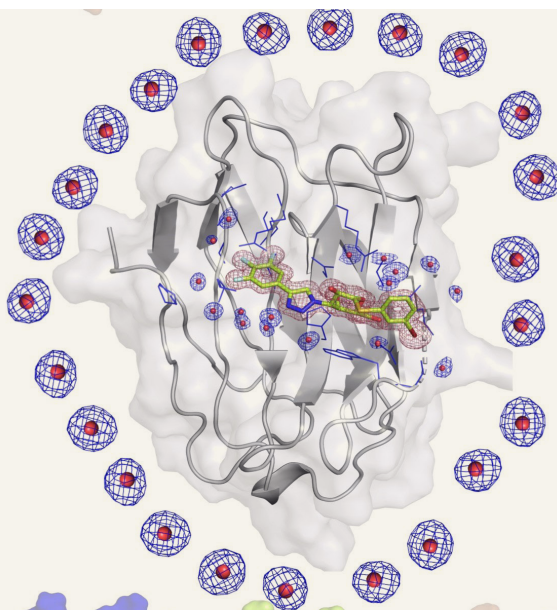
PO Box 117  
221 00 Lund  
+46 46-222 00 00



# Structure and energetics of molecular recognition in galectin-3 ligand interactions

ROHIT KUMAR

BIOCHEMISTRY AND STRUCTURAL BIOLOGY | LUND UNIVERSITY





# Structure and energetics of molecular recognition in galectin-3-ligand interactions

Rohit Kumar



**LUND**  
UNIVERSITY

DOCTORAL DISSERTATION

by due permission of the Faculty of Science, Lund University, Sweden.  
To be defended at KC:B, Kemicentrum, 19th September 2019 at 13:15.

*Faculty opponent*  
Prof. Andreas Heine



<b>Organization</b> LUND UNIVERSITY Biochemistry and Structural Biology P.O. Box 124 SE-22100 Lund, Sweden Author <b>Rohit Kumar</b>		<b>Document name</b> <b>Doctoral Dissertation</b> <b>Date of issue</b> <b>2019-07-05</b> Sponsoring organization	
<b>Title and subtitle: Structure and Energetics of Molecular Recognition in Galectin-3 ligand Interactions</b>			
<b>Abstract</b> <p>Molecular recognition is the key aspect of any cellular and biological function. Two or more molecules interacting with each other cause the effects that drives various basic functions that are fundamental to cells. Be these protein-protein, protein-nucleic acid or protein-ligand interactions, they all play important roles in a cell. Protein-ligand interactions are the most studied as they can have huge implications in not only understanding the basic mechanism of protein function but also for drug design.</p> <p>Protein-ligand interactions are governed by several types of weak non-covalent interaction and the thermodynamics associated with these interactions. What weak interactions the ligand makes with the binding site in protein and how the protein possibly changes conformation to bind ligands are the key to understanding the mechanism involved. One needs structural, thermodynamic and functional data to obtain a complete picture of binding.</p> <p>To study such interactions one needs a protein target, which is galectin-3 in this case. Galectin-3 is a very well characterised member of the medically important galectin family. These proteins bind galactose based carbohydrates to exert their function. Their roles in various cellular functions like apoptosis, differentiation, cell-signaling, cell-cell adhesion and immune responses are well documented. They have been implicated in diseases like tumor formation, metastasis and cardiovascular disease which is unsurprising given their involvement in key cellular functions. The need to study their interaction with ligands is an important step towards understanding their function as well as developing drugs against them.</p> <p>Several methods are used to study these interactions: X-ray crystallography provides an atomic view of the binding, fluorescence polarization provides the binding affinity, isothermal titration calorimetry provides the enthalpic and entropic contributions, neutron crystallography allows us to see hydrogens and thus the hydrogen-bonds involved. Besides these methods, nuclear magnetic resonance and theoretical studies provide energetics of binding. All these methods have been used in this study to provide a complete picture of molecular recognition involved in galectin-3-ligand interactions. The goal was to study basic protein-ligand interactions that can provide insights into designing high affinity and high selectivity inhibitors in the future.</p>			
<b>Key words:</b> Protein-ligand interactions, molecular recognition, weak bonds, thermodynamics, enthalpy, entropy, galectin-3, carbohydrate recognition domain, x-ray crystallography, fluorescence polarization, isothermal titration calorimetry, neutron crystallography,			
Classification system and/or index terms (if any)			
Supplementary bibliographical information		<b>Language:</b> English	
<b>ISSN and key title</b>		<b>ISBN: 978-91-7422-674-4</b>	
Recipient's notes	<b>Number of pages 95</b>		Price
	Security classification		

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature 

Date 2019-07-05

# Structure and energetics of molecular recognition in galectin-3-ligand interactions

Rohit Kumar



**LUND**  
UNIVERSITY

Cover photo by Rohit Kumar

Copyright pp 1-96 (Rohit Kumar)

Paper 1 © Royal Society of Chemistry

Paper 2 © ChemMedChem, Wiley Online Library

Paper 3 © by the Authors (Manuscript unpublished)

Paper 4 © by the Authors (Manuscript unpublished)

Paper 5 © American Chemical Society

Paper 6 © by the Authors (Manuscript unpublished)

Paper 7 © Crystallography Journals Online (IUCR)

Paper 8 © by the Authors (Manuscript unpublished)

Paper 9 © by the Authors (Manuscript unpublished)

Faculty of Science  
Biochemistry and Structural Biology

ISBN 978-91-7422-674-4

Printed in Sweden by Media-Tryck, Lund University  
Lund 2019



Media-Tryck is an environmentally  
certified and ISO 14001:2015 certified  
provider of printed material.  
Read more about our environmental  
work at [www.mediatryck.lu.se](http://www.mediatryck.lu.se)

**MADE IN SWEDEN** 

*To my Family and Friends*

# Table of Contents

Abbreviations .....	8
List of Papers.....	9
List of papers not included in thesis .....	10
Contributions .....	11
Acknowledgement.....	12
Popular science summary .....	16
Summary .....	19
<b>1. Molecular recognition in protein-ligand interactions .....</b>	<b>21</b>
1.1. Interactions governing protein-ligand binding .....	21
1.2. Thermodynamics of protein-ligand binding .....	22
1.2.1. Enthalpic ( $\Delta H$ ) and entropic ( $\Delta S$ ) components .....	23
1.2.2. Solvation and desolvation: Role of waters .....	24
1.3. Non-covalent interactions .....	25
1.3.1. van der Waals interactions.....	25
1.3.2. Hydrophobic interactions .....	25
1.3.3. Hydrogen bonds .....	27
1.3.4. Electrostatic interactions .....	28
1.3.5. Cation- $\pi$ interactions .....	29
1.3.6. Amide- $\pi$ stacking .....	29
1.3.7. Halogen interactions.....	30
1.3.8. Orthogonal multipolar fluorine interactions .....	31
1.4. X-ray crystallography as a powerful tool for drug design .....	32
1.4.1. Brief history.....	32
1.4.2. Synchrotrons.....	33
1.4.3. Basic theory .....	34
1.4.4. Structure based drug design (SBDD) .....	39
<b>2. Galectins .....</b>	<b>41</b>
2.1. Carbohydrate recognition domain (CRD) .....	42
2.2. Ligand binding and valency .....	43
2.3. Galectin functions.....	43
2.3.1. Intracellular and extracellular functions .....	44
<b>3. Galectin-3 .....</b>	<b>45</b>
3.1. Galectin-3 structure .....	45
3.2. Cellular ligands and valency.....	46
3.3. Functions .....	47

3.3.1.	Intracellular functions.....	47
3.3.2.	Extracellular functions.....	47
<b>4.</b>	<b>Aims of the thesis .....</b>	<b>49</b>
<b>5.</b>	<b>Methods .....</b>	<b>51</b>
5.1.	Cloning .....	51
5.2.	Recombinant protein purification .....	51
5.3.	Crystallisation.....	52
5.4.	Crystal manipulations .....	52
5.4.1.	Micro-seeding.....	52
5.4.2.	Macro-seeding and feeding .....	53
5.4.3.	Soaking.....	53
5.5.	X-ray Crystallography .....	53
5.5.1.	Synchrotrons.....	54
5.5.2.	Data collection.....	54
5.5.3.	Data Processing .....	55
5.5.4.	Refinement and model building .....	56
5.6.	Neutron crystallography .....	57
5.6.1.	Deuterated protein production and crystallization.....	58
5.6.2.	Data collection and data processing .....	58
5.6.3.	Joint X-ray/neutron refinement .....	59
5.7.	Other methods used in papers.....	59
<b>6.</b>	<b>Results and Discussion .....</b>	<b>61</b>
6.1.	Paper I.....	62
6.2.	Paper II .....	64
6.3.	Paper III .....	66
6.4.	Paper IV.....	68
6.5.	Paper V .....	70
6.6.	Paper VI.....	71
6.7.	Paper VII .....	73
6.8.	Paper VIII .....	74
6.9.	Paper IX.....	76
6.10.	Results not yet included in manuscripts .....	79
6.10.1.	Neutron data .....	79
<b>7.</b>	<b>References .....</b>	<b>87</b>

# Abbreviations

## *Amino Acid codes*

Glycine	G	Gly
Alanine	A	Ala
Leucine	L	Leu
Methionone	M	Met
Phenylalanine	F	Phe
Tryptophan	W	Try
Lysine	K	Lys
Glutamine	Q	Gln
Glutamic Acid	E	Glu
Serine	S	Ser
Proline	P	Pro
Valine	V	Val
Isoleucine	I	Iso
Cysteine	C	Cys
Tyrosine	Y	Try
Histidine	H	His
Arginine	R	Arg
Asparagine	N	Asn
Aspartic Acid	D	Asp
Threonine	T	Thr

## *Units*

Angstrom	Å (0.1 nm)
degrees celsius	°C
Kelvin	K
kcal	kilocalories
millimolar	mM
micromolar	μM
nanomolar	nM



## List of Papers

- I. **Kumar R\***, Peterson K, Misini Ignjatović M, Leffler H, Ryde U, Nilsson UJ, Logan DT. Substituted polyfluoroaryl interactions with an arginine side chain in galectin-3 are governed by steric-, desolvation and electronic conjugation effects. **Org Biomol Chem.**2019 Jan 31;17(5):1081-1089. doi:10.1039/c8ob02888e.
- II. **Kumar R\***, Misini Ignjatović M, Peterson K, Olsson M, Leffler H, Ryde U, Nilsson UJ, Logan DT. Structure and energetics of ligand–fluorine interactions with galectin-3 backbone and side-chain amides – insight into solvation effects and multipolar interactions. **Accepted ChemMedChem**
- III. Verteramo ML, **Kumar R\***, Misini Ignjatović M, Wallerstein J, Chadimová V, Zetterberg F, Leffler H, Logan DT, Ryde U, Akke M, Nilsson UJ. Structural and thermodynamic studies on halogen-bond interactions in ligand–galectin-3 complexes: Electrostatics, solvation and entropy effects. **Manuscript**
- IV. Wallerstein J, Misini Ignjatović M, **Kumar R°**, Caldararu C, Peterson K, Leffler H, Nilsson UJ, Logan DT, Ryde U, Akke M. Entropy–Entropy Compensation Between the Conformational and Solvent Degrees of Freedom Fine-tunes Affinity in Ligand Binding to Galectin-3C. **Manuscript**
- V. Peterson K, **Kumar R**, Stenström O, Verma P, Verma PR, Håkansson M, Kahl-Knutsson B, Zetterberg F, Leffler H, Akke M, Logan DT, Nilsson UJ. Systematic Tuning of Fluoro-galectin-3 Interactions Provides Thiodigalactoside Derivatives with Single-Digit nM Affinity and High Selectivity. **J Med Chem.** 2018 Feb 8;61(3):1164-1175. doi: 10.1021/acs.jmedchem.7b01626. Epub 2018 Jan 11.
- VI. Noresson AL, **Kumar R\***, Stegmayr J, Carlsson M, Oredsson S, Logan DT, Leffler H, Nilsson UJ. A non-permeable high-affinity sulfated ligand for selective extra-cellular galectin-3 inhibition. **Manuscript**
- VII. Manzoni F, Saraboji K, Sprenger J, **Kumar R**, Noresson AL, Nilsson UJ, Leffler H, Fisher SZ, Schrader TE, Ostermann A, Coates L, Blakeley MP, Oksanen E, Logan DT. perdeuteration, crystallization, data collection and comparison of five neutron diffraction data sets of complexes of human galectin-3C. **Acta Crystallogr D Struct Biol.** 2016 Nov 1;72(Pt 11):1194-1202. Epub 2016 Oct 28.

- VIII. **Kumar R**, Peterson K, Leffler H, Nilsson UJ, Logan DT. Structural perspective of Arg144 mutants of Galectin-3 CRD on ligand binding: role of cation-pi interactions. **Manuscript**
- IX. **Kumar R\***, Mahanti\*. M, Leffler H, Logan DT, Nilsson UJ Ligand sulfur oxidation states stepwise alter ligand-galectin-3 complex conformations. **Manuscript**

\* Shared first author; ° Shared second author

## List of papers not included in thesis

- I. Johansson R, Jonna VR, **Kumar R**, Nayeri N, Lundin D, Sjöberg BM, Hofer A, Logan DT. Structural Mechanism of Allosteric Activity Regulation in a Ribonucleotide Reductase with Double ATP Cones. *Structure*. 2016 Jun 7;24(6):906-17. doi: 10.1016/j.str.2016.03.025. Epub 2016 Apr 28. Erratum in: *Structure*. 2016 Aug 2;24(8):1432-1434.
- II. Caldararu O, **Kumar R**, Oksanen E, Logan DT, Ryde U. Are crystallographic B-factors suitable for calculating protein conformational entropy? **Manuscript**
- III. Johansson R, Aurelius O, Bågenholm V, **Kumar R**, Mulliez E, Logan DT. A structural investigation of allosteric substrate specificity regulation in the class III ribonucleotide reductase from *Thermotoga maritima*. **Manuscript**

# Contributions

## **Paper I:**

I performed the crystallization, data collection and data processing of protein-ligand complexes. I deposited the PDB's. I analyzed the data, made figures and wrote the crystallography part of manuscript, while also contributing to other parts of paper

## **Paper II:**

I performed the crystallization, data collection and data processing of protein-ligand complexes. I deposited the PDB's. I analyzed the data, made figures for the structural analysis. I drafted the manuscript and wrote the introduction part.

## **Papers IV-VI, IX**

I performed the crystallization, data collection and data processing of protein-ligand complexes. I deposited the data in the PDB. I analyzed the data, made figures for the structural analysis

## **Paper VII:**

I performed deuterium labeling of the protein by growing the bacterial cells in minimal media and produced fully deuterated protein for the crystallographic studies.

## **Paper VIII:**

I performed the cloning, expression and purification of mutants. I performed the crystallization, data collection and data processing of protein-ligand complexes. I analyzed the data, made figures for the structural analysis. I wrote the manuscript.

## Acknowledgement

This has been a long and tough, albeit a wonderful journey. I have many people to thank to who have helped me and supported me all this time.

First and foremost, I owe my acknowledgement to my supervisor **Derek Logan**, who has been a great mentor and has helped in shaping my career. Thanks for giving me the opportunity to work in your lab, it has been a real pleasure. You have been really supportive throughout and always encouraged me to learn and pursue new methods. You always provided scientific freedom which really helped me grow as a researcher. You were supportive of my decisions to go for a course or a conference, I greatly appreciate it. Whenever I needed your help or suggestions, you were available. Especially these last days of my writing, you were tremendously helpful. You always believed in me even when I made mistakes. Your positive criticism of my work has helped me improve over the year. You possess sharp insights for research and you always had great ideas. I have learnt a lot from you, be it experiments or writing or scientific thinking. Thanks again Derek, I hope I will get more opportunities in future to work with you again.

I would like to thank **Ulf Nilsson**, my co-supervisor. You always had great ideas/plans for me, discussions with you about the projects were very insightful. You possess immense knowledge in your scientific field and I always tried to learn from you. While taking the medicinal chemistry course I realized you are a great teacher as well. It was a great experience working with you. **Esko Oksanen**, I am really lucky to have worked with you. You have been really supportive and I have learnt a lot about neutron work from you. I admire your intelligence, meetings with you were always great. I really appreciated your help and guidance during my PhD and also your suggestion for my future.

I would like to thank **Mikael Akke**, for great scientific discussions and suggestions during meetings. Also thanks for leading this DecRec project successfully and organising yearly get together, it was fun. I would like to thank **Hakon Leffler**, for his scintillating scientific insights and great stories. I am extremely lucky to have worked with you. I am thankful to **Ulf Ryde** for his critical comments and help with all the manuscripts. I would like to thank other members of DecRec project, **Barbro** for rapid FP data, **Kristoffer**, **Ann Louise**, **Mukul**, **Maria Luisa** and **Alex** for providing me compounds for my work. Thanks to other members of Ulf Nilsson's group. **Majda** and **Octav**, thanks a lot for the collaboration, it was great working with you. **Olof** it has been great knowing you and it was fun to work with you. Too bad we only had one paper together, maybe more in future. **Johan**, thanks for your help and suggestions. We had really long and fruitful project meetings, I really appreciate your methodical approach towards work.

Dear **Francesco**, I really miss you. I would like to express my gratitude towards you, I have learned so many things from you, scientific and extracurricular. I am very happy that I got to travel with you, you were a great companion. And you were a wonderful friend. **Renzo**, it was great working with you, I learned a lot from you. You are very knowledgeable and you have a keen eye for troubleshooting experiments. You were my second mentor in the lab. I really appreciated your help over the years. You are not only a good colleague but a good friend as well. Thanks for all the espresso times we had together.

**Oskar**, it has been my pleasure to know you. You have been a superb colleague and good friend. I remember my first day in lab, you first took me around the department and then you took me walking around Lund and helped me with important paperwork. I am still grateful for that. Over the years you have always kept contact, I really appreciate that. I hope it continues. **Hedda**, my best lab-mate and my best friend. Thanks for a wonderful time when you were in Lund. It was amazing to work with you. I really cherish our trips to Grenoble and Hamburg. I am happy to have a friend like you.

I would like to thanks other lab members past and present. **Wael** thanks for your help in the beginning of my PhD. Thanks to my master students **Niloofar** and **Ragnar**, It was a pleasure supervising you. **Ipsita**, thanks for Bengali sweets and all the best for your PhD.

I had wonderful colleagues and friends at CMPS and I would like to thank them all. **Jennifer**, thanks for a wonderful time at the department and helping me with many things, be it experiments or conferences or any information I needed. It was fun teaching with you. **Stefan**, thanks for collecting my x-ray data ;- ) and thanks for your wonderful camel jokes. Thanks for starting the beer club! **Samuel**, you have been really great. It was a great experience teaching with you. **Mathias**, it has been great knowing you, thanks for being a great colleague. **CJ**, it has been wonderful to know you and talk to you. Thanks for great time at beer club. **Veronika N, Sven, Viktoria, Rebecca, Veronika L** thanks for being good colleagues and for good time at beer club. **Abhishek, Bhakat, Dev** and **Mandar** it has been great knowing you guys, always fun to talk to you all. **Kalyani**, you have been a great colleague and friend, thanks for all the help and suggestions during teaching and wonderful Indian food. Thanks to all previous members of CMPS, **Janina**, thanks for all the scientific discussions, and thanks for a great trip to ESRF. **Alak** you are really great person; it was great knowing you. **Eva**, it has been wonderful to know you. **Tanja** it was great learning experience for me when teaching with you for the first time, Thanks for that. **Diana**, thanks for fun times and great food, I am glad I met you. I would like to thanks **Ingemar, Sarah** and **Susanna** for organizing a wonderful course, it was a great learning experience. **Magnus** thanks for all your technical help and support and for good times at beer club. **Adine. Birgitta** Thanks for the help and

support. **Maryam**, it has been great to know you. Thanks for the talks and thanks for inviting us home. **Per Kjellbom** for providing a good working environment in the department and departmental outings. **Cecilia** thanks a lot for sorting out my prolongation, you are a kind person. **Henrik**, Thanks for giving me the teaching duties. Thanks to all other members of CMPS, it has been great to work here.

Before coming to Lund, I was working in India and I would like to express my deepest gratitude to **Neel Sarovar Bhavesh**, you are wonderful human being and great scientist. This would not have been possible without your help. I am forever grateful to you. I wish to work with you in future.

**RP Singh** sir, thanks for being a wonderful supervisor and thanks for helping me out when I needed it most. I am really grateful to you. It was an excellent experience to do my master's project with you.

**Isha**, thanks a lot for your help and information about coming to Sweden. You are a good senior. When I moved to Lund, I didn't know anyone. **Prashant**, you have been an amazing person and a great friend. I really appreciated your care and help. You were the greatest room-mate. It was a fun experience. We also worked together, hopefully we can publish it someday.

I have made some wonderful friends over the year in Lund. **Jennifer** thanks for introducing me to your friends, who became my friends as well. **Emmy**, we became good friends over a really short period, which was surprising given I take a lot of time to make friends. Now we are a family, hopefully it stays that way. **Neha** it has been my pleasure to know you, I think Emmy couldnot have made a better choice ;-). Thanks for bearing me over all these years. **Nimba, Hafsa and Rishi** thanks for good times. Thanks everyone for all the great weekends and weekdays. Thanks for the Christmas fun and secret santa. **Sunny, Dhanu** thanks for always inviting me, it has been great knowing you guys. Dhanu thanks for amazing biriyani. I have loved the time spent with **Hansika and Ishank**. **Kailash, Nida, Vivek, Vidit and Sameeksha**, it has been a great experience interacting with you guys. Thanks for all the wonderful times during Christmas vacations. Thanks for inviting me to your home all of you. **Manish and Snigdha** thanks for the good times and good food.

I would express my gratitude to my great friends in India. They have been great support to me. Cheers to **Saurabh, Birendra, Purushottam, Awinash and Paritosh**. It has been a great journey and you guys have been awesome. I am really lucky to have friends like you. **Ritu and Pragyan** thanks for the good times, it has been wonderful knowing you both. **Kuldeep**, thanks for helping me and giving suggestions during tough times. You have been a great senior and Friend. **Prashant**, you have been a great friend, we went through tough times together, I value your friendship. **Nabila**, I am really grateful to have a friend like you, thanks for

everything. **Rashmi**, you have been a wonderful person, thanks for all the help and support throughout.

**Preksha** thanks a lot for being there as a friend during my tough times. **Sneha** and **Karthik** thanks for your care I really cherish your friendship.

I cannot thank enough my family. They have been my source of inspiration and strength. **Maa** and **Papa**, you guys always believed in me and made sure I get the best education. Whatever I have achieved the credit goes to my parents. Thanks to my wonderful sisters **Swati** and **Missi**. Thanks for lovely nieces **Pihu**, **Nimmy** and **babu**, they are the most adorable people to me. Thanks to **Ritesh** Ji and **Nishikant** Ji for being part of my family. all my relatives in India. Thanks to my In-laws for supporting us. Thanks to **Mummy**, **Papa**, **Ranjeet** and **Sanjeet** Bhaiya and **Sonam** Bhabhi and **Pari** Bhabhi. Thanks dear **Monika** for great times in Delhi and for supporting me and Soni during tough times. Finally, I would like to convey all my gratitude and love to **Soni**. I have known you for 13 wonderful years. You have been my immense support. I could not have been successful without you. You are the source of positive energy I my life. You make every day amazing and I look forward to many such years.



## Popular science summary

Cells are the smallest structural, functional and biological units of all multicellular living organisms like us, humans. A human body has approximately trillion cells that perform their functions to help in growth, immunity and survival. Cells are composed of four different types of macromolecules (individual working units): nucleic acids, lipids, carbohydrates and proteins. Nucleic acids constitute the genome (DNA) which carries genetic information, and RNA, which has a wide variety of functions. Lipids form the cell membranes and act as structural unit. Carbohydrates, also called polysaccharides, are made of sugar residues like glucose, galactose etc. They act as an energy source, play role in signalling when bound to proteins and can form cell walls. Proteins are the most versatile among all the macromolecules, and they are the workforce of the cell. A human cell can have atleast 20 000 types of proteins depending on its function. They perform diverse roles in cells, like forming structural elements, e.g. the cytoskeleton (microtubules), they are the catalysts of the cell and perform all the enzymatic processes, they act as transporters and channels that help in movement of molecules in and out of cells and they act as signalling/communication molecules by carrying information between/within cells. Proteins are able to perform these functions because they interact with the other macromolecules and numerous small molecules called ligands. These protein-ligand interactions are very important for cellular functions and the basic understanding of how the protein functions relies on studying such molecular recognition.

Proteins are made up of amino acids arranged in specific orders (decided by the genetic code) and folded into a three-dimensional functional unit. Each protein has different combination of amino acids, which decides its structure, location and function. Most proteins are found at specific locations like the cytoplasm, nucleus or extracellular space and their expression is cell specific. Dysregulation of protein expression and function can lead to imbalances in cellular processes like cell division, cell death and cell differentiation. That can develop into diseases, hence studying structure and function of proteins is important, and protein-ligand interactions are the fundamental aspect to these studies.

All protein-ligand interactions are chemical in nature, as proteins are essentially big, organized chemicals. Thus the interactions they make with other molecules are purely chemical. They involve several weak non-covalent bonds (like hydrogen bonds, hydrophobic interactions, electrostatic interactions etc.) between protein and ligand atoms. The type and number of weak interactions define the thermodynamics and affinity of binding. Hence one needs to understand the intricacies of these chemical interactions to clearly elucidate the mechanism of binding.

To achieve that end, specific tools are needed and there are various methods available to perform the desired study. Proteins are very small molecules; their size is defined in units called Ångströms (Å) which is 10<sup>-10</sup> meters and an average protein has a size of about 100 Å. Ligands are even smaller units. So how do we see such small molecules? X-ray crystallography is one powerful method that allows us to see the small molecules, it basically takes an atomic picture of them. X-ray crystallography can give a complete atomic structure of the molecules and we are able to see all the details of binding and the weak interactions involved.

A drawback to X-ray crystallography is that we cannot see hydrogens, and so we cannot easily study hydrogen bonding between ligands, solvent (water) and protein. To address this, neutron crystallography is used, in this method the protein is labelled with deuterium (<sup>2</sup>H) which is an isotope of hydrogen but has one extra neutron. The data from neutron crystallography allows us to see all the hydrogens(deuterium) and the hydrogen bonding.

Neutron and X-ray crystallography are great complimentary tools to study mechanisms of binding. However, this information is not sufficient alone, as the interactions are not rigid, but are quite dynamic processes involving conformational changes happening in both protein and ligand. Thus, we need more information like binding affinity, thermodynamics (enthalpy and entropic contributions) and conformational changes. Isothermal titration calorimetry (ITC) is an excellent technique to find the affinity as well as thermodynamics of binding. Nuclear magnetic resonance (NMR) is another powerful method to study conformational changes and the entropy associated with protein-ligand binding. Theoretical studies are also a very useful tool to understand the energetics of binding. These studies are performed by complex simulation and calculations.

We have the tools and we have the idea, so next we need is a target, a model on which to perform our experiments. In this thesis the target molecule is galectin-3, a member of the galectin family of proteins that bind galactose-containing saccharides on the surfaces of proteins. Galectin-3 is found everywhere in the cell as well as on the extracellular surface. Galectin-3 interacts with a plethora of proteins and exert its function. Galectin-3 is involved in various cellular processes like cell division, immune response, cell death and cell differentiation. Thus one can assume if they start misbehaving there will be bad implications. That is what it has been observed: galectin-3 is involved in various diseases like cancer, immune disorders and cardiovascular diseases. There has been lot of studies on this protein in order to understand its mechanism of action,

Next we need ligands to perform binding studies, and they are synthesized from complex organic reactions. To begin with galactose was chosen as the basic scaffold as galectin-3 binds galactose. Several modifications were made and they were studied using previously mentioned methods. This gave further insights into what

changes to make, what atoms to use for substitution. So, it was a feedback process, where new compounds were synthesized based on data accumulated from old ligands.

To summarize, in this thesis we have used galectin-3 as the model to study basic aspects of molecular recognition using tools like X-ray and neutron crystallography, ITC, NMR and theoretical studies. The findings have been reported in several papers, some of them has been published.

## Summary

In recent times, structure-based drug design has been the most preferred method to find new drugs against possible or well-known drug targets. The method relies on atomic structures of target-ligand complexes. X-ray crystallography has been at the core of this revolution: no other structural methods give us the possibility to study protein-ligand interactions and drug design at the atomic level. Each protein has a defined expression level and any alteration can lead to differences in their activity. Most diseases involve differential expression of particular proteins, hence making them a drug target.

In this thesis we have tried to study basic protein-ligand interactions while also contributing to find a high affinity inhibitor. The target is from a family of proteins called galectins. They bind to  $\beta$ -galactoside-containing ligands. They are found everywhere in the cell and they are involved in several key processes involving signalling, apoptosis, and other immunogenic pathways. Their expression becomes unregulated during diseases like cancer, cardiovascular disorders and their roles in these diseases have been very well documented. My main target was galectin-3, which is unique among galectins in having a N-terminal repetitive domain and a highly conserved C-terminal carbohydrate recognition domain (CRD). We worked mostly with the CRD as this domain recognises the carbohydrates. The structure of the galectin-3 CRD has been reported previously in complex with ligands based on  $\beta$ -galactoside.

Protein-ligand interactions are fundamental and important for cells to function and for their capacity to interact with their surroundings. These are complex chemical and physical processes. Multiple weak non-covalent bonds and the thermodynamics associated with them are the driving forces for protein-ligand interactions. However, conformational entropy of protein is equally important, how the protein changes conformation upon binding a ligand is key to such interactions. The ideal binding involves several non-covalent bonds that compensate for the loss of ordered water, and the enthalpy and entropy of the complex itself drives the binding to be favourable or unfavourable.

In the current work we have focused on modifying the ligands at key places, like introducing fluorines and also changing the position of these substitutions. Then we determined structures of all these complexes to see the effect of the substitutions on the binding. Binding affinities were determined to give an idea how these subtle changes affect the protein-ligand interactions thermodynamically. The results add not only to our current understanding of basic protein-ligand interactions but also provide insights into making very high affinity and selective ligands against galectins.



# 1. Molecular recognition in protein-ligand interactions

Proteins are the most versatile molecules present in living cells. They have several critical roles, ranging from structure, signalling and catalysis to reproduction. They help cells in carrying out almost all of the necessary tasks to function and survive. Proteins are found everywhere in the cell, from the nucleus to cell membranes and the extracellular matrix. Some are needed for maintaining structure, some are on the cell surface and receive signals from the surroundings, while some are enzymes and perform catalysis to generate energy or products needed for survival. Proteins achieve their function by interacting with other molecules (proteins, lipids, sugar and nucleic acids). The interaction usually involves binding a ligand, which can be any small molecule like a sugar, an amino acid, a nucleic acid etc. These interactions, which can be transient or long-lived, are mediated mainly by noncovalent interactions, although there are examples where covalent bonding is involved as well<sup>1-3</sup>. These interactions are highly specific for each protein-ligand complex. This, the protein-ligand interaction is the key to study the many basic functions of proteins in general and it can give us insights into drug design as well. Hence studying protein-ligand interactions is paramount to deciphering the basics of molecular recognition as well as having a tremendous application in fighting diseases<sup>4</sup>.

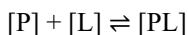
## 1.1. Interactions governing protein-ligand binding

Proteins are highly specific for the kind of ligand they bind and the affinities also vary from the millimolar to the picomolar range depending on how well the ligand binds in the binding pocket. This affinity and specificity of ligand binding to their target proteins are controlled to some extent by weak noncovalent interactions, which are much weaker than covalent interactions. The strength of a covalent single bond is usually in the region 80-100 kcal/mol, but non-covalent interactions are much weaker, usually in the range of 1-3 kcal/mol. Hence several weak interactions are required for successful binding of ligands, thus providing differences in affinity and specificity. Apart from the weak interactions, other crucial factors that can affect affinity and specificity are thermodynamic (enthalpic,

entropic and desolvation) parameters. This means that water molecules are crucial in the binding affinity and that the weak interactions affect thermodynamic parameters. Steric clashes are another important parameter that can affect the affinities greatly. Binding affinity is a function of two quantities, the binding enthalpy and the binding entropy, which are affected by these weak interactions. Studying these weak interactions is essential for rational drug design and lead optimization<sup>5</sup>. We will look at these factors in following sections.

## 1.2. Thermodynamics of protein-ligand binding

Protein-ligand interaction is a highly flexible process, and binding is governed by the energetic contributions of the noncovalent interactions and the dynamics of the ligand and solvent. The precise elucidation of molecular recognition processes involving protein-ligand interactions requires a complete characterization of the binding energetics and correlation of thermodynamic data with the interacting structures involved. Binding affinity is the strength of the reversible interaction between the protein and ligand. It can be described as dissociation constant  $K_d$ :



$$K_d = \frac{[P][L]}{[PL]}$$

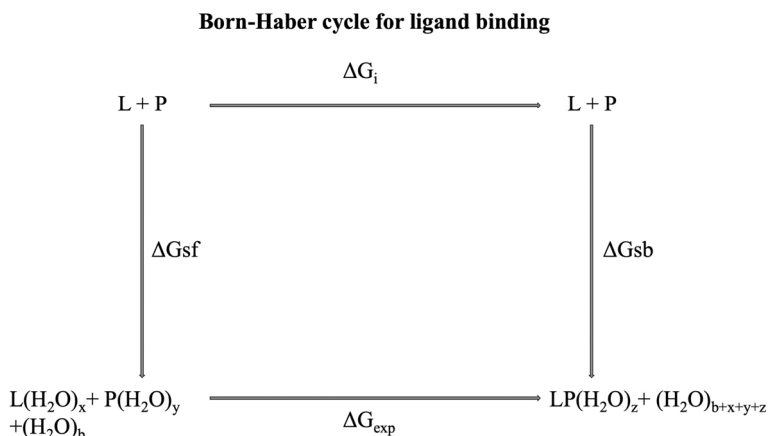
Where  $[P]$ ,  $[L]$  and  $[PL]$  are protein, ligand and protein-ligand complex concentrations respectively. A quantitative description of the factors that govern molecular associations requires determination of changes of all thermodynamic parameters, including free energy of binding ( $\Delta G$ ), which depends on the enthalpy ( $\Delta H$ ), and entropy ( $\Delta S$ ) of binding<sup>6,7</sup>. Like any other spontaneous process, a noncovalent binding event takes place only when it is associated with a negative binding free energy ( $\Delta G$ ), which is the well-known sum of an enthalpic term ( $\Delta H$ ) and an entropic term ( $-T\Delta S$ ). These terms may be of equal or opposite sign and thus lead to various thermodynamic signatures of a binding event, ranging from enthalpic to entropy-driven<sup>8</sup>. A detailed analysis of many protein ligand complexes show that majority of the interactions are enthalpically favoured<sup>8</sup>. The Gibbs free energy change ( $\Delta G$ ) of binding is the most important thermodynamic description of the event, since it determines the stability of any given protein-ligand complex, and it has been the greatest analytical tool for the characterization of structure–function relationships. Isothermal titration calorimetry (ITC) is the best method to study the thermodynamics of binding and gives quantitative measurements of binding enthalpy and entropy<sup>6</sup>.

$$\Delta G^\circ = -RT \ln K_d$$



$$\Delta G = \Delta H - T\Delta S$$

$\Delta G^\circ$  is the standard binding free energy, R is the universal gas constant (1.987 cal K<sup>-1</sup> mol<sup>-1</sup>) and T is the temperature in K. It can be deduced that binding affinity of a ligand is not only dependent on the precise spatial disposition of interacting groups and their contributions to intermolecular interactions but also by the dynamics of these groups. Thus, consideration of both these factors is of utmost importance in determining the correct binding affinity. Since entropic and enthalpic components of binding are highly dependent on many system-specific properties, the practitioner has to conclude that optimizing for free energy is still the only viable approach to structure-based design<sup>9,7</sup>.



**Figure 1:** A Born-Haber representation of ligand binding<sup>9-11</sup>.  $\Delta G_{\text{exp}}$ ,  $\Delta G_{\text{sf}}$ ,  $\Delta G_{\text{sb}}$ , are the experimental free energy, solvation free energy for free ligand and protein and protein-ligand complex respectively. P and L are protein and ligand respectively.  $\Delta G_i$  is the intrinsic free energy of binding. They are related by the following equation:

$$\Delta G_{\text{exp}} = \Delta G_i + \Delta G_{\text{sb}} - \Delta G_{\text{sf}}$$

### 1.2.1. Enthalpic ( $\Delta H$ ) and entropic ( $\Delta S$ ) components

In simple words,  $\Delta H$  represents the changes in noncovalent bond energy occurring during the interaction. The enthalpy change of binding reflects the loss of protein-solvent hydrogen bonds, formation of protein-ligand bonds, salt bridges and other weak contacts, and solvent reorganization near protein surfaces. These individual components may produce either favourable or unfavourable contributions. The dissection of each noncovalent interaction is not feasible since the net heat effect of a particular bond is the balance between the reaction enthalpy of the ligand to the protein and to the solvent. Also, structural alterations at the binding site due to the binding event may contribute to the binding enthalpy<sup>6,12</sup>.

Binding entropy represents one of the major driving forces. The main factor contributing to  $\Delta S$  of complex formation is solvation and desolvation effects. Since the entropy of solvation of polar and hydrophobic groups is large, the burial of water-accessible surface area on binding results in solvent release which often makes a large and positive contribution to the total entropy of interaction. Another important, unfavourable contribution reflects the reduction of rotational degrees of freedom around the torsion angles of protein and ligand side-chains<sup>6</sup>. Entropy is commonly regarded as a problem for larger and more flexible ligands, since presumably more conformational degrees of freedom would be lost upon binding.<sup>13</sup>

A widely observed feature of protein-ligand binding thermodynamics is the seeming tug-of-war between enthalpy and entropy. In general, enthalpic interactions improve selectivity due to their geometric specificity, and they are inherently more efficient since they tend to be larger in magnitude than entropic effects<sup>13,14</sup>. Protein-ligand complexes that exhibit more negative enthalpies of binding, mostly do so at the cost of more positive  $-T\Delta S$  terms and *vice versa*. This enthalpy-entropy compensation effect is clearly evidenced in protein-ligand complexes. The ligands with the most favourable entropies of binding actually have positive enthalpies of binding. Conversely, the most favourable enthalpies have very unfavourable entropies of binding<sup>13</sup>. Rigid ligands reduce the entropic penalty and improve affinity. Although recent studies have shown that most flexible ligands have the entropically most favored binding. These flexible ligands trap water molecules more efficiently prior to binding and release them after binding, resulting in favourable entropic binding<sup>15</sup>.

### 1.2.2. Solvation and desolvation: Role of waters

Water plays a crucial role in protein structure and function, while also governing the ligand binding. Water is unique as a solvent: it can act as both H-bond donor and acceptor and can bridge H-bonds between atoms. However, their contribution to ligand binding is not only limited to H-bonding: they have major implications for the thermodynamics of binding. Structured water molecules in ligand binding sites are crucial, and replacing them with ligands (desolvation) favours increase of entropy, which can lead to a thermodynamically favourable process, depending on its relationship to enthalpy<sup>16-18</sup>. Binding is a two-step process involving desolvation and association<sup>19</sup>. Thus, if the energy required to desolvate both ligand and binding pocket is greater than the stabilisation energy gained by binding, then the ligand may be ineffective<sup>18,20,21</sup>. Increased stabilization of water molecules results in enthalpically more favourable binding and enhanced affinity<sup>22</sup>. Also, there is an enthalpic penalty for removing water molecules that are important for the protein. Therefore, careful analysis and design are necessary to fully utilise the water network. The desolvation process is crucial to other intermolecular interactions like hydrophobic interactions, electrostatic interactions (salt bridges) and halogen

bonding. These are all governed by water displacement and increase in entropy<sup>19,11</sup>. It is possible to gain direct information on the role of solvent molecules in protein ligand complexes using X-ray crystallography, as solvent molecules that are in fixed positions relative to the protein can often be identified. Paper III in this thesis addresses the importance of desolvation and role of waters in the affinity of binding in context of galectin-3 inhibitors.

## 1.3. Non-covalent interactions

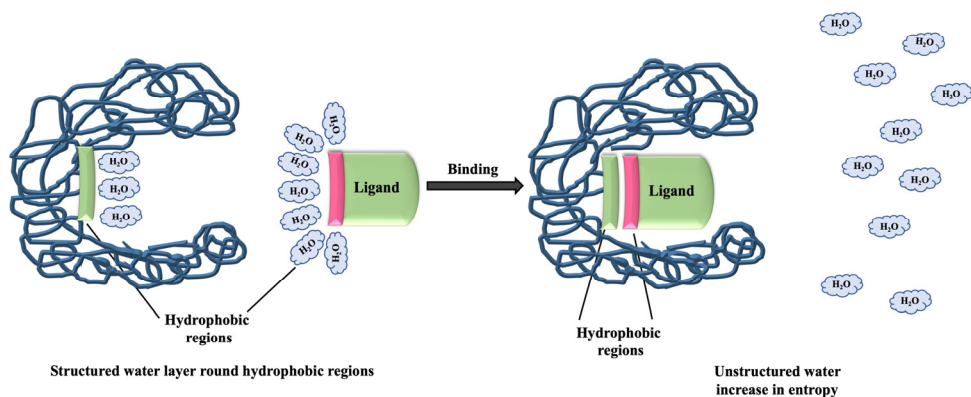
In 2017 Schapira et al. performed a systematic analysis of 11000 protein-ligand complexes to find out what weak interactions are involved and which weak interactions are most prevalent. They found seven most common interactions<sup>23,24</sup>, namely hydrophobic, hydrogen bonds and  $\pi$ - $\pi$  stacking, although other less frequent interactions like cation- $\pi$ , amide- $\pi$  and halogen bonds are also quite important in protein-ligand interactions.

### 1.3.1. van der Waals interactions

Van der Waals interactions are a general and broad category of relatively weak interaction (0.5-1 kcal/mol) and are non-ionic in nature. They happen when two atoms come close enough and their electron clouds touch, resulting in charge dispersion and formation of weak dipoles. They are general atomic interactions and are always present between atoms if they are close enough. They are non-specific and nondirectional, so we will focus on only specific interactions governing protein-ligand binding.

### 1.3.2. Hydrophobic interactions

These are most common interactions in protein-ligand complexes<sup>23</sup>. These interactions are based on the hydrophobic effect, which is basically an energetically favourable process that brings hydrophobic groups together in aqueous solutions by displacing water molecules. The single best structural parameter correlating with binding affinity is the amount of hydrophobic surface buried upon ligand binding. On the magnitude of the hydrophobic effect was estimated to be around 0.7 kcal/mol, or a 3.5-fold increase in binding constant for a methyl group<sup>25</sup>. These interactions dominate the free energy of protein-ligand binding and are pivotal to protein-ligand

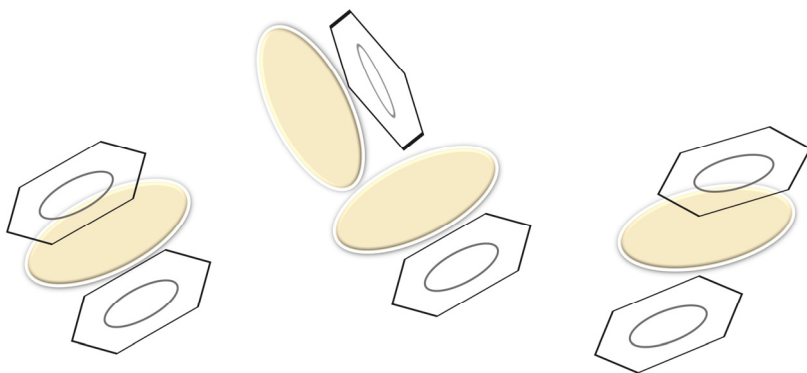


**Figure 2:** Diagrammatic representation of hydrophobic interaction and role of water

recognition and drug design<sup>5</sup>. Several types of hydrophobic interactions occur when a ligand binds a protein, like aromatic-aromatic, aliphatic-aromatic, aliphatic-aliphatic etc. Hydrophobic interactions add mostly to the entropy of ligand binding. Proteins have several hydrophobic residues that add to the hydrophobic interactions, like Val, Leu, Ile, Phe, Trp, Tyr etc. Ligands have alkyl groups and aromatic groups as well. Other interactions like the sulfur atom in Met with aromatic groups in ligands are also prevalent.

#### *Aryl-Aryl interaction/ $\pi$ -Stacking interactions*

This is the most common form of hydrophobic interaction<sup>23</sup>. Most drug candidates have at least one aromatic ring. These groups can form  $\pi$ - $\pi$  stacking interaction with other hydrophobic residues in the protein (Trp, Phe and Tyr). These interactions can be edge to face or face to face. This  $\pi$  stacking can also happen between pyranose rings in ligands and aromatic amino acids in proteins, which is common for sugar binding proteins.

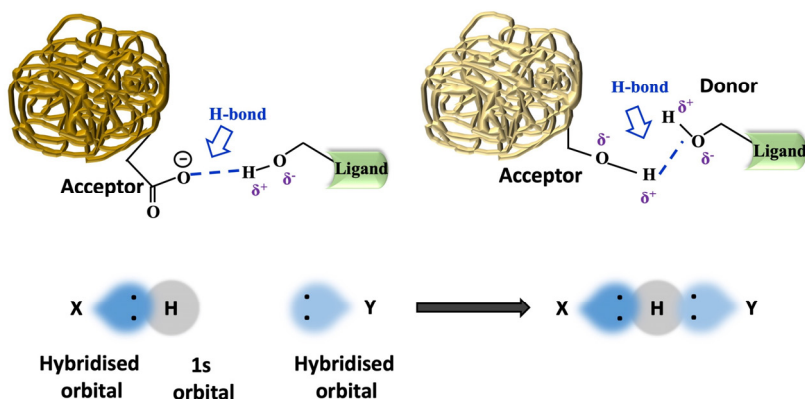


**Figure 3:** Aryl-aryl stacking interactions shown in cartoon form. Face to face, edge to face and edge to edge stacking is shown.  $\pi$  system is represented by yellow disc.

### 1.3.3. Hydrogen bonds

These are the second most frequent interactions in protein-ligand complexes and one of the most important ones<sup>23,9</sup>. These bonds have a donor and an acceptor. Donors are N, O and F and the acceptor is an electronegative atom with lone pair of electrons. They are denoted as Donor-H...Acceptor, where Donor is covalently bound to H. These interactions are directional and distance-dependent and are the predominant contributor to molecular recognition and specificity. There are multiple hydrogen bonds involved in binding of ligands, so they are the most important for specificity. The energy of a hydrogen bond can be between 0.2–40 kcal/mol<sup>26</sup>, and the donor-acceptor distance is less than 3.5 Å<sup>23</sup>.

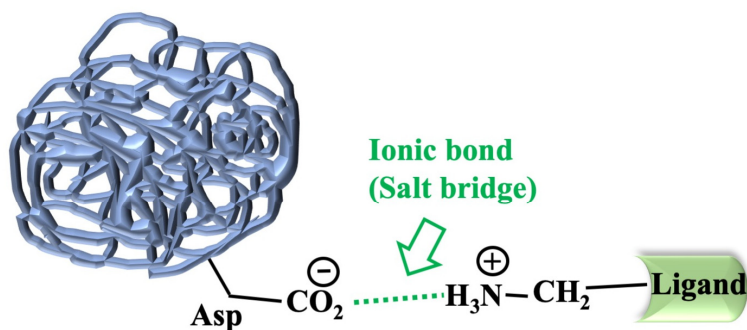
N-H...O interactions are most common, followed by O-H...O and N-H...N (ref. <sup>23</sup>). These are classified as strong hydrogen bonds and the distances are close to 3 Å, whereas weak hydrogen bonds involving C atoms, mostly C-H...O, have a distance of around 3.5 Å. The angular preference of these bonds are quite pronounced, the angle for Donor-H...Acceptor is generally above 150°. The direction of hydrogen atoms and angles define specificity. Given the prevalent role of H-bonds in ligand affinity and specificity, as well as the importance of their geometry and directionality, their structural elucidation becomes necessary. However, X-ray crystallography is mostly unable to see hydrogens, so that leads us to use neutron crystallography where we are able to see the hydrogens clearly.



**Figure 4:** Pictorial representation of H-bond between donors and acceptors. The orbital representation shows the sharing of electrons.

### 1.3.4. Electrostatic interactions

This is a type of charge-based interaction involving two oppositely-charged groups. Electrostatic interactions always involve charged residues, in proteins (Asp, Glu, Lys, Arg, His) and in ligands it could be phosphate, sulfate, amine etc. This interaction is mostly between a positively charged nitrogen and negatively charged oxygen, and either could be from protein or ligand. Arg acts as the cation in most of these interactions<sup>23</sup>. The binding energy gained from forming a salt bridge is not always sufficient to compensate for the energetic penalty of desolvating charged groups<sup>27</sup>. However, the strength of these interactions relies heavily on the environment.

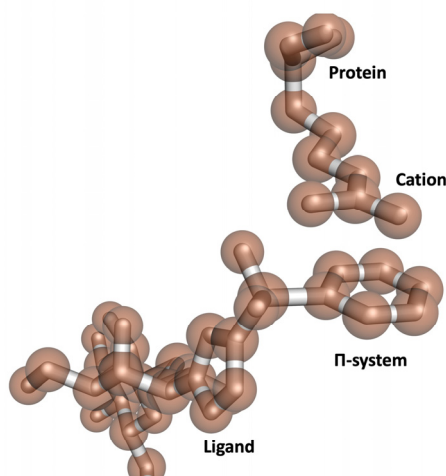


**Figure 5:** Cartoon representation of salt bridges between charged groups of protein and ligand

### 1.3.5. Cation- $\pi$ interactions

These have been extensively studied in protein structures, especially in galectin-ligand structures. These interactions occur between a positively charged group and an electron-rich aromatic group. The cation can be a positively charged N in Arg/Lys, with Arg being the most favoured one. An analysis of structures deposited in the Protein Data Bank (PDB) showed that the aromatic ring is almost always from a ligand and the cation is from a protein. These interactions can be described as a type of electrostatic interaction. The free energy for these interactions is around 5 kcal/mol<sup>28</sup>.

Significant cation-  $\pi$  interactions are rarely buried and prefer to be exposed to the solvent. Engineering a surface exposed cation-  $\pi$  interaction can drastically improve protein stability and affinity of ligand binding<sup>28,29,30</sup>.



**Figure 6:** Cation- $\pi$  interaction depicted in this image, where cation is a charged residue and the  $\pi$ -system is an aromatic group from a ligand

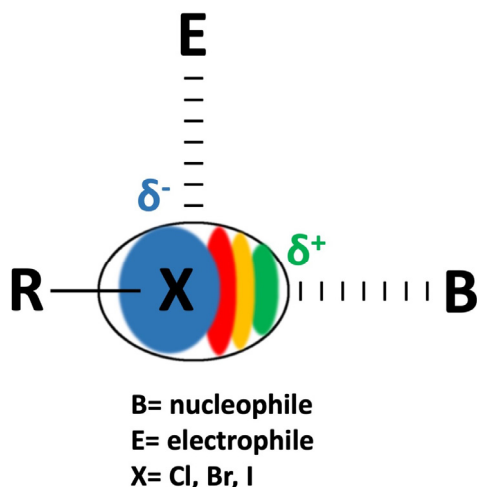
### 1.3.6. Amide- $\pi$ stacking

These interactions are canonical to aromatic  $\pi$  stacking interaction, where the  $\pi$ -surface of an amide bond stacks against the  $\pi$ -surface of an aromatic group. These can be also face to face or edge to face kind of interactions, although there is no preference to one like in aromatic  $\pi$  stacking interactions<sup>23</sup>. Most common amides were from Gly and Trp for face-to-face, whereas Gly and Leu form edge to face stacking interactions with the ligand aromatic groups<sup>23</sup>.



### 1.3.7. Halogen interactions

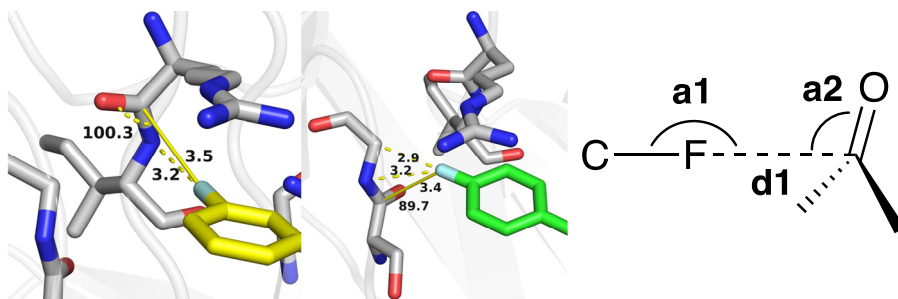
These are one of the most important interactions found in recent studies on inhibitor design. Halogens are introduced into ligands as a strategy to enhance affinity and selectivity. The halogen X(Cl, Br, I) in a C-X bond can interact with electrophiles, nucleophiles, waters and other halogens<sup>31,32</sup>. Fluorines are highly electronegative and the least polarizable of the halogens, whereas the heavier halogens have unique electronic properties when bound to aryl or alkyl groups. They show an anisotropy of electron density with a positive area of electrostatic potential on the halogen opposite to the C-X bond<sup>9,31</sup> called the  $\sigma$ -hole. Halogen bonding occurs between the positive electrostatic potential of a covalently bonded halogen atom that acts as a Lewis acid and an electron rich atom (N, O and S) that acts as a Lewis base. The halogen atom acts as electron acceptor (halogen bond donor) and the electron rich atom acts as electron donor (halogen bond acceptor)<sup>32</sup>. In biological systems, halogen atoms not only form short C-X $\cdots$ O-Y interactions with the protein (O-Y is a carbonyl, hydroxyl, charged carboxylate, or phosphate group), but can also accept hydrogen bonds from hydroxyl groups or water molecules and form halogen-water-hydrogen (XWH) bridges, C-X $\cdots$ H-O. The X $\cdots$ O distance is shorter than or equal to sum of the van der Waals radii (3.27 Å for Cl $\cdots$ O, 3.37 Å for Br $\cdots$ O and 3.50 Å for I $\cdots$ O)<sup>31</sup>. Strength of halogen bond increases with the size of halogen, although these interactions are significantly weaker than hydrogen bonds, their energies are similar to weak H-bonds (2.5 kcal/mol) and they have lower desolvation cost<sup>9</sup>. Halogen bonding has been the main focus for paper V, where we have established their role in improving binding and specificity for galectin ligands.



**Figure 7:** Representation of halogen bond  $\sigma$ -hole represented in green colour which is positively charged.

### 1.3.8. Orthogonal multipolar fluorine interactions

These are multipolar interactions between a halogen (mostly fluorine) and an electrophilic group like the amide bond in protein backbone and side chain. Fluorine is a small (van der Waals radius 1.47 Å) but highly electronegative atom and can easily be used as a substitute for hydrogen (van der Waals radius 1.20 Å) without causing steric clashes. Recent studies have shown that introducing a fluorine at key positions in ligands enhances affinity, beside affecting physicochemical properties of ligands in a positive way like improved metabolism and solubility<sup>33</sup>. Distinct fluorophilic groups in proteins include the abundant peptide bonds, which form multipolar C-F...H-N, C-F...C=O, and C-F...H-C $\alpha$  interactions, as well as the side-chain amide groups of Asn and Glu and the positively charged guanidinium group of Arg<sup>34,33</sup>. A “fluorine scan” conducted for a class of highly preorganized inhibitors of thrombin has helped to identify favourable interactions of organofluorine such as orthogonal dipolar interactions with backbone C=O residues<sup>33</sup>. Unlike the head-to-head interactions in halogen bonds, the interactions happen more or less orthogonally with the carbonyl groups<sup>23</sup>. Fluorines can interact with both polar and hydrophobic groups in proteins<sup>35,36</sup>. C-F unit is a poor H-bond acceptor as organic fluorine has very low proton affinity and is weakly polarizable. C-F...H-N (backbone amide) interactions are energetically favourable hence abundant in protein-ligand structures, and the distance between F and N is approximately 3.5 Å. But the main interactions are orthogonal multipolar C-F...C=O interactions.



**Figure 8:** Fluorine atoms interacting with the carbonyl groups in gal3gal3 CRD. The angles a1 and a2 are distance d1-dependent

The C-F bond is generally inclined to the F-C axis with angles adopting values typically between 100°–160°, but rarely 180°, at short to middle-range contact distances<sup>36,37</sup>. The combination of X-ray crystal-structure analysis of a protein–ligand complex, small-molecule X-ray crystallography and database mining has for the first time shown that H-C $\alpha$ ...C=O fragments provide a pronounced fluorophilic environment. The high frequency of H-C $\alpha$ ...C=O units in the active sites of proteins suggests that such F interactions could be effectively exploited for

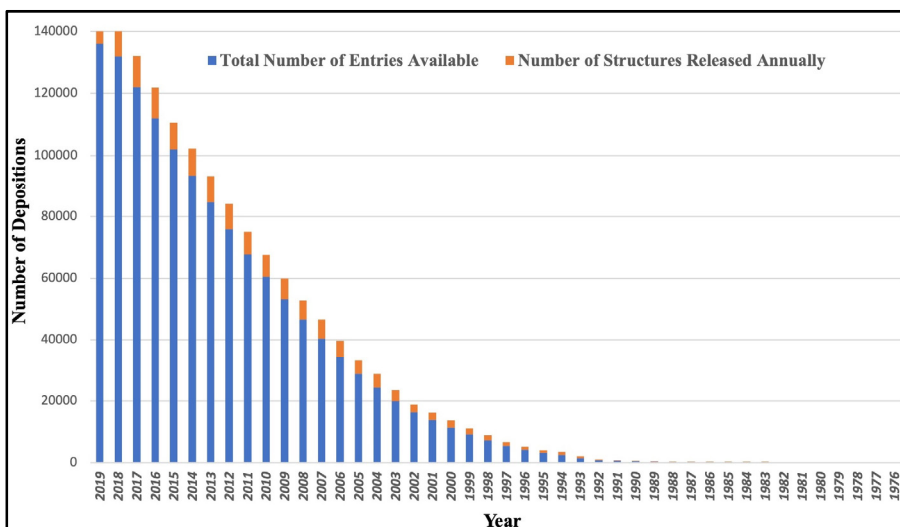
enhancing ligand affinity or selectivity in structure-based inhibitor design<sup>37</sup>. At shorter F $\cdots$ C=O distances ( $< 3.0$  Å), the F $\cdots$ C=O angle tends towards 90°, while at longer distances, the angular dependence is weaker. The C–F $\cdots$ C=O angle is more variable<sup>33</sup>. In apolar environments, the orthogonal multipolar fluorine–amide interaction with backbone amides can contribute  $-0.2$  to  $-0.35$  kcal/mol in binding free energy ( $\Delta\Delta G$ )<sup>33</sup>. These interactions are the main focus of papers II and IV in my thesis, where we have shown that number and position of fluorines can be a strategy in drug design, although other factors like desolvation and energetics play a role as well.

## 1.4. X-ray crystallography as a powerful tool for drug design

Macromolecular X-ray crystallography has been at the forefront of modern rational drug design. Recent advances in crystallography, like tuneable X-ray beams, synchrotrons and automation of structure solution have further revolutionized the field. This has been the only method for a long time to provide an atomic view of protein-ligand complexes. Although NMR and cryo-EM have improved with time, crystallography is still the method of choice for drug design. The method involves producing crystals of proteins/protein-ligand complexes and then exposing them to X-rays. The resulting diffraction pattern is recorded and then an electron density map is generated through complex mathematical calculations (Fourier transform). The better the resolution, the more detailed is the map. Then the protein/ligand model is optimised to fit the data in a process called model building and refinement.

### 1.4.1. Brief history

Crystallography is a fairly old method: the first protein to be crystallized was haemoglobin in 1851 by Funke<sup>38</sup>. X-rays were discovered in 1895 by W.C. Roentgen who was awarded the Nobel Prize for this discovery in 1901. In 1912, Max von Laue did the first diffraction studies using X-rays. W.L. Bragg did a diffraction studies on a chemical crystal in 1913 to obtain its structure and postulated Bragg's law of diffraction, which is still used to solve the diffraction data<sup>39</sup>. Pepsin crystals were the first made to diffract X-rays by Northrop in 1930. The first protein structure ever to be determined was that of myoglobin in 1958 by Kendrew<sup>40</sup>.



**Figure 9:** The yearly growth in structures solved by X-ray crystallography. The first breakthrough was recombinant protein expression that happened around early 1980's. The second major revolution happened in the mid-1990s when synchrotrons began to be used for protein crystallography. (Source: Protein Data Bank)

## 1.4.2. Synchrotrons

In the early days of crystallography, data were collected with home X-ray sources that were small machines that generated low intensity X-rays. Data collection typically took days. It was a low throughput process. The synchrotrons arrived in 1950s, but started to be used for crystallography in the 1970s. Synchrotrons produced X-rays that were very intense and data collection was reduced to a few hours. A synchrotron is an extremely powerful source of X-rays. The X-rays are produced by high energy electrons as they circulate around the storage ring. A synchrotron accelerates electrons to extremely high energy and when they change direction periodically in a magnetic field they lose energy in the form of radiation. The resulting X-rays are emitted as well-collimated beams, each directed toward a beamline next to the accelerator. The X-ray beams emitted by the electrons are directed toward beamlines where the experiments are carried out. Further improvements over the years have seen tremendous improvements in the brilliance of the X-ray beams and the possibility to tune the energy and wavelength of the radiation to suit the experiments. The beams have become more brilliant and one data collection takes now only a few seconds. It is now possible to collect hundreds of data sets in one day, which is of immense help in the field of drug discovery. Other improvements that have made remarkable impact are better detectors, software, and automation like using robots to mount crystals. I have collected all my data at synchrotrons like the ESRF (France), DESY (Hamburg, Germany) and

the MAX IV Laboratory (Lund, Sweden). MAX IV is a next-generation synchrotron and it produces the most brilliant X-rays in the world<sup>41</sup>.

### 1.4.3. Basic theory

Protein crystallography involves three distinct steps. First you need to produce crystals from your target protein, the second step is to collect data, usually at a synchrotron, by exposing the crystals to X-rays and the final step involves data processing, solving the phase problem, refinement and model building. The first two steps require a lot of lab work to get good data and the final step requires a lot of computer work to get a complete atomic model. The detailed methodology used for my work has been included in methods section. We will focus more on the principle behind the method, the results and its implications in rational drug design. For detailed theory one can refer to excellent crystallography books such as those from Gale Rhodes<sup>42</sup>, Bernhard Rupp<sup>43</sup>, Tom Blundell and Louise Johnson<sup>44</sup>.

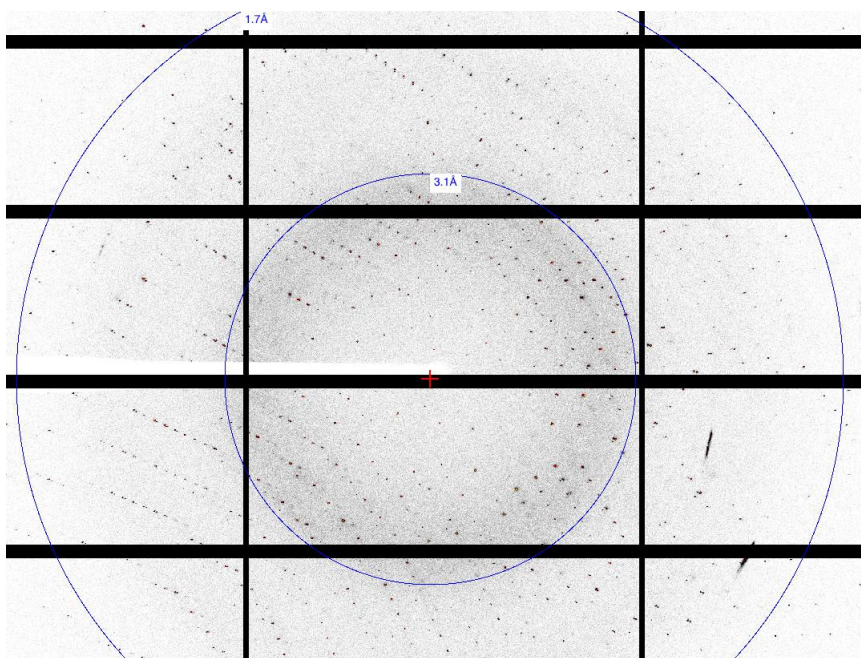
#### *Crystallization*

Crystals are repetitive ordered arrays of molecules in three dimensions governed by symmetry. A unit cell is the smallest unit of volume that contains all the structural and symmetry information and its repetitive translation along the principal axes can generate the whole crystal. The asymmetric unit is the smallest unit of volume that contains the structural information. A unit cell can have one or more asymmetric units, and the asymmetric unit can have one or more protein molecules. The protein crystals have voids even when tightly packed because of their irregular shapes, and these voids are occupied by disordered solvent (mostly water). Getting a good crystal is major bottleneck. To get the structure of a target protein one needs to express and purify the protein in large amounts, ideally 10 mg or more. The purified protein is then subjected to vapour diffusion methods, where the protein is mixed with a precipitant solution and is allowed to equilibrate with the same precipitant solution in a well so that the precipitant concentration becomes similar to the well, to get well-diffracting crystals. One often has to screen hundreds of conditions to get the crystals, which are then further optimized, and this is a time-consuming process. For drug design purposes ideally we choose a target that is already crystallized, but one can always decide on a new drug target. As the conditions for crystallization are already known, the next task is to incorporate the ligands in the crystals, which can be achieved either by soaking the apo protein crystals in ligand solution or by pre-mixing the protein with ligands and then crystallizing them together.

### *Data collection*

When data are collected at a synchrotron the crystals are typically cooled to 100K to avoid radiation damage<sup>45</sup>, although room temperature data collection is possible as well. Crystals are exposed to X-rays of the desired energy and the diffraction pattern is recorded on a detector. There has been tremendous development in detectors, starting from film to CCD detectors and then pixel-based detectors that are highly sensitive, have fast readout time and can record very small signals. The fast readout time has helped in making the data collection shutterless, which means the shutter can remain open while the crystal is being rotated and data collection is going on<sup>46,47</sup>. This has dramatically improved the data collection time and data quality. The crystal is rotated along an axis that is perpendicular to the x-ray beam, and the reason is to collect multiple parts of different reciprocal planes. As the crystal is rotated other set of planes are exposed to x-ray and we see new diffraction spots; this is how full coverage of reciprocal space is obtained. The rotation range for collecting a complete data is dependent on the crystal space group.

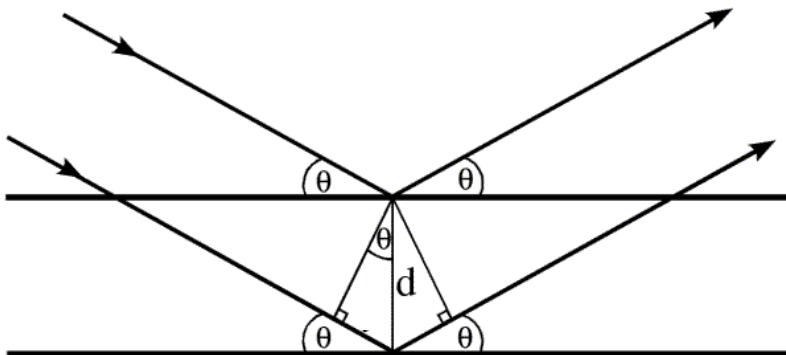
The diffracted beams are recorded, and their intensity of diffraction depends on the crystal packing. The spacing between spots indicates the size of the unit cell. The larger the unit cell the closer the spots and vice-versa. Data collection needs a strategy like exposure time, number of images, and degree of rotation. These factors are experiment- as well as crystal dependent and depend on the diffraction pattern.



**Figure 10:** A diffraction pattern for a galectin-3 CRD crystal recorded on a Pilatus detector.

### *Diffraction and data processing*

Crystals diffract X-rays and the intensity of diffraction can be recorded. The diffraction is governed by Bragg's Law as shown in the equation. It assumes crystals are made of lattice planes and the diffraction is governed by the interplanar distance and angle of incidence.



**Figure 11:** Diffraction by lattice planes separated by distance  $d$  with incident and diffracted ray at an angle  $\theta$

When the diffracted waves are in phase, they produce constructive interference, and these reflections are related by following equation:

$$n\lambda = 2d\sin\theta$$

$\theta$  is the angle of incident and reflected X-rays,  $\lambda$  is the wavelength of X-rays used,  $d$  is the distance between the planes and  $n$  is an integer/order of reflection. This law also defines the resolution of the data and the maximum resolution achievable is  $\lambda/2$  when  $\theta = 90^\circ$ .

Why do we need X-rays to obtain the structures of molecules? The reason is that X-rays have wavelengths that are in the range of atomic distances ( $\text{\AA}$ ) and hence can provide accurate information about the atoms and bonds. X-rays interact in an elastic manner with electrons in the molecule and are scattered. This means the wavelength is unchanged after scattering.

Each spot on the detector is called a reflection and comes from interaction of X-rays with atoms in the crystal unit cell. The so-called structure factor is calculated from these spots. The structure factor  $F_{hkl}$  is a mathematical function that describes amplitude and phase of a diffracted wave from a set of lattice planes characterized by Miller indices  $(h,k,l)$ . The intensity  $I_{hkl}$  of reflections is proportional to the square of the structure factor  $F_{hkl}$ . It can be represented as a summation of waves scattered from every atom in the unit cell by the following equation:

$$F_{hkl} = \sum_{j=1}^n f_j e^{2\pi i[hx + ky + lz]}$$

where  $f_j$  is the atomic scattering factor,  $n$  is the number of atoms in unit cell, and  $xyz$  are the positional co-ordinates of the atoms. The electron density can be derived from structure factors using a Fourier transform (FT). The FT relates functions in mutually reciprocal domains (in crystallography, real and reciprocal space) in unique and invertible form. So structure factors can be transformed into electron density and vice -versa. Electron density  $\rho(xyz)$  can be represented by following equation:

$$\rho(xyz) = \frac{1}{V} \sum_{hkl} |F_{hkl}| e^{-2\pi i[hx + ky + lz]} - \phi(hkl)$$

where  $V$  is the volume of unit cell. The structure factors from known atomic structure can be represented by vectors in an Argand Diagram (amplitudes and phases) and their numerical values in an Argand diagram. The phase ( $\phi$ ) information is missing from the experimental data and that creates a phase problem which needs to be solved in order to get the structure. The phase problem is generally solved by using a homologous structure as template (called molecular replacement) or by phasing methods using heavy atoms or anomalous scattering using intrinsic atoms in protein like sulfur. The phases are very important as they carry most of the information, so they dominate the maps and are more important than amplitudes.

Once the data is collected it is processed to get information about space group and unit cell dimensions. The data reduction and processing consist of several steps. The first step is to index the strong diffraction spots and deduce unit cell constants, space group, crystal orientation and mosaicity information. Next step is to integrate all the images while also refining crystal orientation and detector parameters. In mathematical terms we are trying to measure intensities of diffraction spots to obtain structure factor amplitudes. The final step is to “reduce” the data by merging and scaling the intensities of multiple observations of a reflection. The data processing statistics are critically assessed to make sure data is of good quality, completeness is acceptable, there is no radiation damage and the resolution has been determined correctly. There are several statistical terms that can give this information like  $CC_{1/2}$ <sup>48</sup>,  $R_{meas}$ ,  $I/\sigma$  (signal versus noise)<sup>49</sup> etc. To get a good electron density map, accurate data processing is indispensable.

### *Refinement and model building*

Once the electron density map and the model are ready, the model is iteratively refined to make the model fit to the experimental data better. In the process the map and phases are improved and atomic co-ordinates are adjusted to fit the diffraction data. The model is further build to correct for any stereochemical anomaly and to



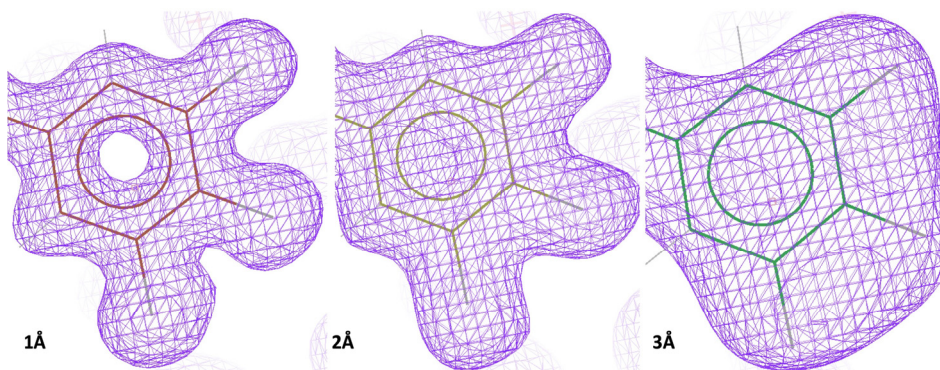
better fit the electron density. The overall fit between diffraction data and model is represented by a statistical term called R value between the scaled structure factor amplitudes  $F_{obs}$  and  $F_{calc}$ .

$$R = \frac{\sum |F_{obs} - F_{calc}|}{\sum F_{obs}}$$

$R_{work}$  and  $R_{free}$  are the two statistical terms that show how well the data is refined and how good the model is.  $R_{free}$  is the R-value of a small subset of data that is kept aside during refinements (generally 55%) to cross-validate the refinement process.

### *Role of resolution*

The resolution of the final model is dependent on how well the crystals diffracted and how well the electron density has been resolved and refined. The resolution of the final model is important for drug design purposes. A high-resolution structure gives more information and hence is ideal for drug design. Ideally one would want a structure of 2.5 Å resolution or higher for SBDD. But the size of the target protein is also relevant, if a protein is big then even a low-resolution structure (3.5 Å) can give some useful insights. To comprehend what the resolution offers, the following images provide a hint:



**Figure 12:** Part of a ligand (aromatic ring with three fluorines) showing the detail and shape of electron density associated with resolution. Better resolution (lower value) leads to unambiguous and distinct electron density. As one can see the low-resolution map (3 Å) is mostly a blob and it will be difficult to tell what group it fits. The phases for each panel are same and are calculated from the same model.

Thus, with higher resolution you get better and accurate information, and the drug design process is more precise. In this thesis, all the structures are of very high resolution (1.0-1.3 Å), which provide most accurate atomic description of all the interactions and precisely distinguishes the moieties in ligands.

#### 1.4.4. Structure based drug design (SBDD)

As the name implies, this kind of drug design is based on the structure of the target protein. The drug design process involves target identification, lead identification, and lead optimization. SBDD forms important part of lead discovery and lead optimization process where the protein structure and ligand scaffold are known, and the focus is to further modify the existing ligand scaffold to achieve thermodynamically favourable and high affinity and selectivity inhibitors. A desirable end result is a successful drug candidate. SBDD has accelerated the process of drug discovery, because the target is known and the binding pocket is well defined so designing of well binding ligands is quite rapid.

X-ray crystallography has been the principal method to achieve this and with the development of synchrotrons and detectors, data collection speed and quality has improved drastically. The quality of the target structure matters: a better and high-resolution structure is ideal for drug design. Once the structure is known and binding site is identified, the search for a lead compound begins. Ideal way is to start with a natural ligand for the target, which is galactose in case of galectins. This is where Lipinski's Rule of Five<sup>50</sup> comes in. This rule states that a good ligand to begin with should be under 500 Da, have less than 5 H-bond donors, less than 10 H-bond acceptors and an octanol-water partition coefficient (logP) of less than 5. Although variations of these rules occur, these are generally good while starting on a ligand candidate.

Then the process of lead optimization begins. The optimization process takes into consideration the binding pocket, the residues involved and the scaffold of the natural ligand, if available. Medicinal chemists use following strategy for lead optimization:

- Vary substituents
- Elongate the scaffold
- Expand/contract rings
- Vary ring type
- Structure simplifications
- Rigidification
- Isosters: moieties having similar size and chemical properties



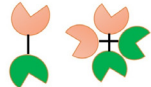
Combining the above information gives a fair idea about the modification of the scaffold, but numerous substitutions, extensions and introduction of different chemical moieties are necessary. All this is governed by the chemical properties of the binding pocket; for example, in a hydrophobic pocket, methylation of the ligand could be good idea. In a similar way, if there are charged residues, introducing

oppositely-charged moieties in the ligand is beneficial. Basically, the goal is to optimize every possible interaction. Then the task is to find the binding affinities of the ligands, which is achieved by some high throughput assay such as fluorescence polarization, which has been used in this thesis. ITC also plays a major part, not only in finding affinities but the enthalpic and entropic contributions. However, ITC is low-throughput and used for selective inhibitors showing potential. This gives more data to work with and helps in identifying ligands that not only bind with high affinity but are also thermodynamically favourable. Lead optimization is initially assisted by computer-aided docking, which gives scores of the ligand fit. Based on good scores, certain ligands are selected and synthesized, then X-ray crystallography and ITC are performed. The X-ray structure gives an exact picture of how the ligand binds and ITC gives information of the affinity and thermodynamics. With the onset of high-throughput screening and fragment-based drug design, it is possible to scan a plethora of drug candidates and analyse them. SBDD has been a really successful approach and it continues to grow. There are numerous drugs on the market that were designed using this approach. Drugs against HIV protease are mostly from SBDD<sup>51,52</sup>. The drugs against thymidylate kinase involved in cancer were one of the first from this method<sup>51</sup>. There is a huge potential for more drugs against G-protein coupled receptors (GPCR). These proteins are involved in several diseases and several conventional drugs on the market are targeted towards them. With advancement in the crystallography and reporting of several GPCR structures, one can hope that more efficient drugs will come soon<sup>51</sup>.

## 2. Galectins

Lectins are a superfamily of proteins that bind to various carbohydrates attached to proteins and lipids. Galectins are a sub-family of soluble lectins that specifically bind to  $\beta$ -galactoside-containing carbohydrates<sup>53,54</sup>. They are highly conserved evolutionarily and are found in all classes of multi-cellular living beings from nematodes to mammals<sup>54</sup>.

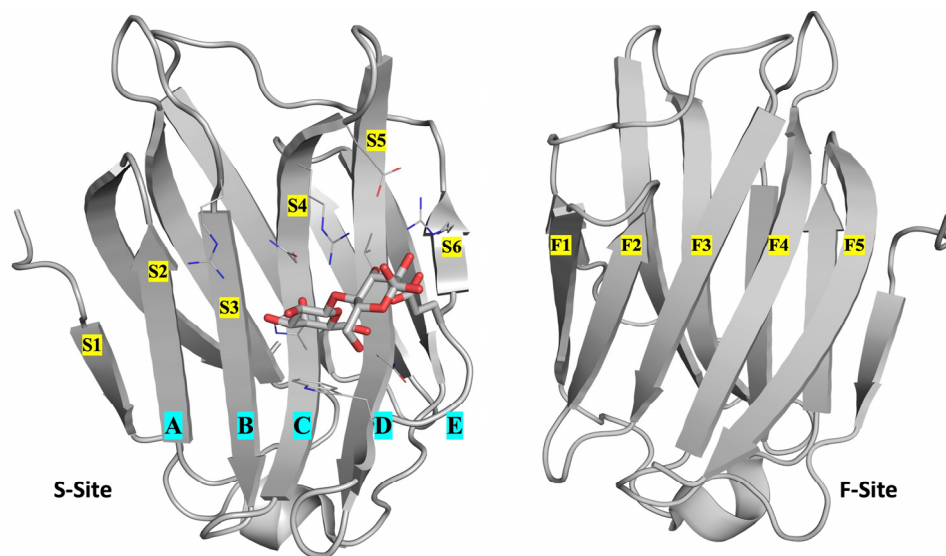
Galectins have a highly conserved carbohydrate recognition domain CRD<sup>53,54</sup>. At present there are 15 different kinds of galectins in mammals<sup>55</sup>. They are divided into three proto-types based on their domain organization. They have either one or two CRDs<sup>54,56</sup>. Type I or prototype galectins include galectin-1, -2, -5, -7, -10, -11, -13, -14 and -15 with only one CRD; type II or chimera type galectins (galectin-3) have one CRD and a N-terminal repetitive domain; type III or tandem-repeat type galectins include galectin-4, -6, -8, -9 and -12. These galectins have two homologous CRDs connected by a linker<sup>55</sup>. Type I can be either monomers or dimers, and each monomer can bind carbohydrates. Type III has two carbohydrate binding-sites formed by two CRDs. Galectins are expressed in all cell types, sometimes at cytosolic concentrations as high as 5  $\mu$ M, although expression varies among cell types<sup>57</sup>. Their expression is highly regulated and tissue-specific. They are synthesized in the cytosol and function mostly in the cytoplasm and nucleus, but they are also secreted extracellularly by a non-classical pathway as they lack a classical secretory signal sequence<sup>58,59</sup>. Galectins perform a plethora of cellular processes, and years of research have established their roles in tumour development and progression, immune and inflammatory responses, neural degeneration, atherosclerosis and diabetes<sup>56</sup>. In this thesis we will mostly discuss human galectins.

Type	Structure	Galectin
Prototype (one CRD)	 monomer dimer	1, 2, 5, 7, 10, 11,13,14,15
Chimeric type		3
Tandem Repeat (Two CRD )	 monomer dimer	4, 6, 8, 9, 12

**Figure 13:** Classification of galectins.

## 2.1. Carbohydrate recognition domain (CRD)

CRDs from galectins are highly conserved among mammals and are composed of about 130 residues<sup>56</sup>. The first structure of a galectin-CRD was solved in 1993 (PDB id 1SLT)<sup>60</sup>, a bovine galectin-1. The first human galectin CRD structure was solved in 1998 (PDB id 1A3K), a galectin-3 CRD<sup>61</sup>. The structure is composed of a sandwich of two  $\beta$ -sheets, one with five and another with six  $\beta$ -strands. This structure is quite conserved in all the galectin structures solved to date. The six-stranded sheet is concave, and the groove called the S-site has the binding pocket for the sugar. The other side is called the F-site on the five  $\beta$ -strands that binds other CRDs and proteins<sup>56,55</sup>. The carbohydrate binding site can accommodate adjacent saccharides as well as galactose<sup>62</sup>. The carbohydrate binding site can be divided into 5 subsites A-E<sup>56</sup>, as represented in Fig. 13. C is the site where galactose binds; the subsites on either side (A-B & D-E) can fit other sugar molecules that are part of an oligosaccharide. Binding of galactose in subsite C involves a highly conserved amino acid sequence. Binding in subsite D is the second most conserved and the structural requirement in this subsite is flexible for the interactions, which can be fulfilled by different disaccharides. This provides the variation in specificity among different galectins. Site E is poorly defined and it binds to moieties attached to the reducing end of the ligand in site D. These moieties could be another saccharide, lipids or proteins<sup>56</sup>. Binding sites for non-carbohydrate ligands have also been identified on the CRD, for example the F-site on the CRD of galectin-8 has specific affinity for NDP52 (nuclear domain 10 protein)<sup>63,64</sup>.



**Figure 14:** S-site showing the 6 strands (S1-S6) and the five binding pocket subsites (A-E). F-site showing 5 strands (F1-F5). This representation image is from galectin-3 CRD.

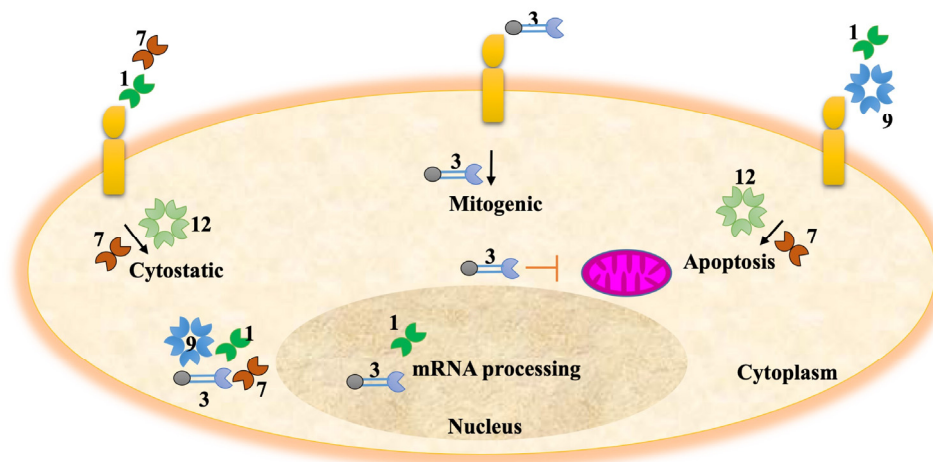
## 2.2. Ligand binding and valency

The functions of galectins are determined by what targets they bind (intracellular or extracellular) and how they bind. Most of the galectins known are either bivalent or multivalent in binding their ligands. They are known to form ordered arrays of protein lattices upon binding their targets<sup>65,66</sup>. Galectin CRDs bind galactose with millimolar affinity, they bind common disaccharides with mid-micromolar affinity and can bind their target glycoconjugates with sub-micromolar affinity<sup>55</sup>. Thus, their binding affinity increases if galactose is attached to other saccharides<sup>67</sup>.

## 2.3. Galectin functions

Galectins have a diverse array of functions in a cell. They are found in intracellular as well as extracellular compartments. Their functions are governed by their localisation: they perform different sets of functions when inside the cell compared to when outside the cell. They are synthesized in the cytosol by ribosomes and have no signal peptides<sup>56,68</sup>. However, some galectins are secreted by the cell through non-classical secretory pathway into extracellular compartments<sup>69</sup> that bypasses the Golgi-ER vesicular transport<sup>59</sup>. These non-classical pathways may involve accumulation of galectins on the cytoplasmic side, which are then either secreted

through exosomes or direct transport through the plasma membrane<sup>70</sup>. Most galectins are found in multiple kinds of tissues while some are restricted to specific tissues. Their expression is regulated in normal tissue. They are involved in the regulation of inflammation and immunity, progression of cancer and cell differentiation<sup>56</sup>. They are unique in many ways: for example, same galectins can have roles in the nucleus as well as regulation of cell adhesion and signalling outside cells<sup>56</sup>. In the following sections, we will focus on important functions of galectins in several cellular processes and their implications in diseases.



**Figure 15:** A pictorial representation of some of the functions associated with all the galectins.

### 2.3.1. Intracellular and extracellular functions

As mentioned earlier, galectins can be found both intracellularly and extracellularly and they have different functions depending on the location. Galectins on the outside are able to interact with cell surface glycoproteins, they can also interact with glycoconjugates in extracellular matrix like laminin, fibronectin etc.<sup>71,72</sup>. They can bind bivalently or multivalently and cross-link the glycoproteins. This can lead to either receptor endocytosis and regulation or a signalling cascade. These have effects on cellular processes such as mitosis, apoptosis and cell-cycle progression<sup>62</sup>.

While inside the cell, galectins can move between cytoplasm and nucleus and are involved in mRNA splicing, apoptosis and cell growth regulation<sup>73</sup>. Galectins interact with several proteins within the cell and these are mostly carbohydrate-independent protein-protein interactions<sup>74,63,64</sup>.

### 3. Galectin-3

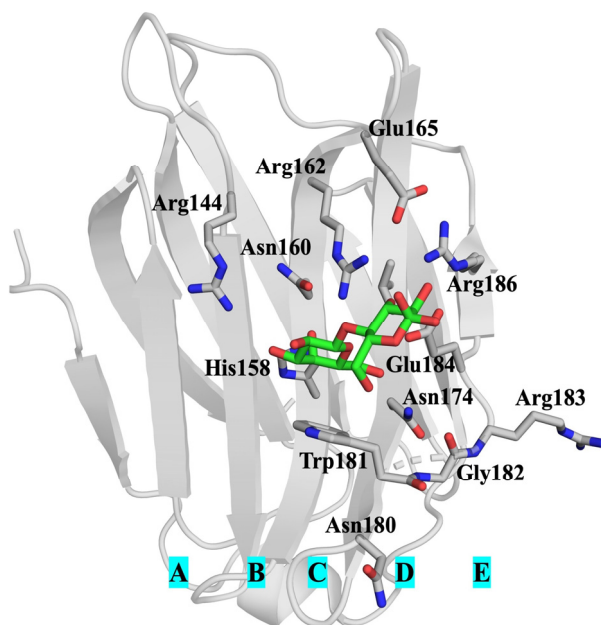
Galectin-3 is the most studied galectin. It is the only member of the chimera type galectins, with a CRD and an N-terminal domain that is involved in oligomerization<sup>66,75,76</sup>. It is coded by the LGALS3 gene in humans. It was first identified as cell surface antigen Mac-2 on macrophages<sup>77</sup>. Galectin-3 is synthesized in the cytosol but is found to function in the nucleus and outside of cells. It performs a plethora of important cell functions because of its diverse set of interactions with both extra and intracellular targets. Its expression is finely tuned and highly regulated<sup>78</sup>. Galectin-3's role in numerous biological processes such as cell–cell and cell–matrix interactions, growth, proliferation, differentiation, and inflammation are well documented. Its involvement in such crucial processes makes it important in several human diseases such as cancer, fibrosis, chronic inflammation and cardiovascular diseases<sup>79–82</sup>.

#### 3.1. Galectin-3 structure

As mentioned, galectin-3 is the only member of chimera type galectins. Apart from the conserved CRD, it has an atypical non-carbohydrate binding N-terminal domain of 100-150 residues<sup>56,79</sup>. This domain helps in the oligomerization of the protein upon binding to carbohydrates and formation of lattices. There is still no clear evidence about the mechanism of oligomerization<sup>75,76</sup>. The N-terminal domain (ND) is highly flexible and is important for biological activity of galectin-3. The ND sequence is also conserved among galectin-3s from different species<sup>83–85</sup>. It contains 7-9 homologous repeats of Pro-Gly-Ala-Tyr-Pro-Gly-X-X-X, but lacks any charged or hydrophobic residues<sup>79</sup>. The sequence of the ND has some similarity with collagen  $\alpha 1$  chain<sup>86</sup>. The first structure of a galectin-3 CRD was solved in 1998 by Seetharaman et. al<sup>61</sup>. Since then, many structures of human galectin-3 CRD have been solved with various ligands, 65 in total to date, including our own contributions. The structure of the CRD is quite similar to that of other galectins and has conserved residues at the key binding site, as discussed in the previous chapter. Here, we will discuss more specific interactions and the residues involved. The main binding site C where galactose sits is lined by an NWGR motif with Asn180, Trp181, Gly182 and Arg183 (and other residues His158, Asn160, Asn174) which



is also found in anti-apoptotic Bcl2 proteins<sup>79,87</sup>. The NWGR motif is not only responsible for self-association of the CRD in the absence of saccharides but also plays a role in the interaction with other proteins<sup>88,89</sup>. The full-length structure of galectin-3 has not been solved yet because of the highly flexible and intrinsically disordered ND, but it has been shown that the ND interacts with the CRD in some cases<sup>76,90–92</sup>. Recently there has been a successful attempt to crystallize parts of the ND separately as well as with the CRD, but still the complete picture is lacking<sup>93</sup>. The ND is also supposed to play a role in non-classical secretion of galectin-3 outside the cell<sup>94</sup>. The ND is susceptible to proteolysis by matrix-metalloproteinases, which regulates their function in extracellular compartment<sup>95</sup>.



**Figure 16:** CRD of galectin-3 showing the binding site with key residues and the binding subsites (A-E) shown. The NWGR motif is at the bottom right of the domain.

### 3.2. Cellular ligands and valency

As we know, the binding pocket in the CRD recognises galactose, although the affinity is in the millimolar range. Introducing a glucose at the reducing end of galactose to make lactose increases the affinity for galectin-3 50-fold<sup>96,97</sup>. The exchange of the hydroxyl group to acetamide group makes for the even higher affinity ligand N-acetyllactosamine (LacNac). Adding another galactose group on 3'-OH to this LacNac further increases the affinity 23 times in comparison to

lactose<sup>97</sup>. With most of its biological ligands, galectin-3 interacts via LacNac residues present on the glycan moieties. However, not all LacNac-containing glycoproteins are galectin-3 targets because of poorer binding affinity compared to other galectins<sup>79</sup>. Galectin-3 possesses bivalent as well as multivalent binding properties despite having only one CRD. The reason is the presence of the ND which multimerizes when the CRD binds its ligand<sup>76</sup>.

### 3.3. Functions

Galectin-3 is found in the cytoplasm, the nucleus and extracellular spaces and interacts with several targets, which influences a plethora of cellular processes. It performs specific functions according to its location. Therefore, it is best to sort these functions by their compartmentalization.

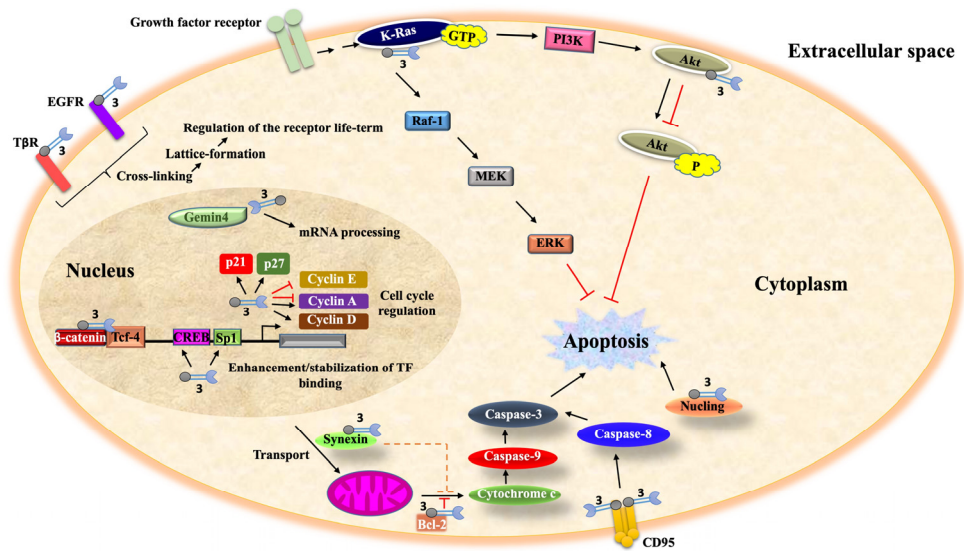
#### 3.3.1. Intracellular functions

Galectin-3 is found both in the cytoplasm and the nucleus<sup>98</sup>. It shuttles through the two compartments depending on the requirements; in some cell types it is found mostly in the cytoplasm whereas in others it is found mostly in the nucleus<sup>83,99–102</sup>. In the cytoplasm it interacts with Bcl-2 and inhibits apoptosis. Galectin-3 acts as an activator of oncogenic K-Ras proteins that promote cell proliferation<sup>103</sup>. They also interact with Akt (protein kinase B) proteins that regulate cell cycle<sup>104</sup>. They interact with synexin in mitochondria and regulate apoptosis<sup>105</sup>.

In the nucleus, it interacts with gemin-4 that is involved in pre-mRNA splicing<sup>73,74,106</sup>. It also interacts with the Wnt pathway protein  $\beta$ -catenin that is involved in inhibition of apoptosis, cell cycle regulation and transcription regulation of several tumour associated genes<sup>107,108</sup>.

#### 3.3.2. Extracellular functions

Galectin-3 binds to several glycosylated membrane proteins like integrins and extracellular matrix components like laminin, fibronectin *etc.* and regulates processes like cell-cell adhesion, immune-regulation, angiogenesis and metastasis<sup>71,79,109–111</sup>. It binds  $\beta$ 1 integrins and regulates their endocytosis, resulting in immune reactions<sup>111</sup>. Based on these findings one can see how important this protein is, which make it a great drug target. Figure 15 gives an overview of its functions.



**Figure 17:** Major functions of galectin-3 (denoted by 3) in the cells based on their location.

## 4. Aims of the thesis

The aim of my thesis was to study protein-ligand interactions using the galectin-3 CRD as the model. Galectin-3, owing to its important cellular functions, presents itself as a wonderful target, as evident from previous sections. Besides, the structure of galectin-3 CRD was already known, so it makes it a great template for studying molecular recognition in protein-ligand interactions as well as SBDD<sup>21,61</sup>. The work published by Saraboji *et. al.*<sup>21</sup> showed the binding pocket in great detail at very high resolution. The work also established the detailed water network in the ligand binding pocket. Based on these structures, several subsites were identified, as described previously. More work by Hakon Leffler and Ulf Nilsson *et al.* showed the binding modes of mono-thiogalactosides and di-thiogalactosides and established how certain substitutions and extensions increased affinity<sup>30,112–115</sup>. In this thesis I have utilized those subsites to design ligands and study the effect of different substitutions on the binding affinity and thermodynamics. I have also utilized the information gathered from previous galectin-3 CRD-ligand complexes to build on the knowledge on how to improve affinity and selectivity by tinkering with the ligand scaffold and the thermodynamics of binding.

This thesis work is a part of a bigger project involving diverse groups from synthesis to biology to structure to theoretical studies. I was involved with the structural part. The primary goal of my thesis was to solve X-ray and neutron structures of different sets of compounds against galectin-3 CRD. The compounds synthesized were varied with one or more substitutions to make several series. For example, in paper I the compound series was varied at only one position, giving subtle changes in binding affinity. Similarly, the compounds in paper II varied in position and amount of fluorination. The compounds were also divided into two categories based on mono- or di-thio-galactosides. Di-thio-galactosides are of higher affinity as they have more weak interactions than the ones with one galactose. The affinities are determined by fluorescence polarization which has been the high-throughput method for this project<sup>116</sup>. The thermodynamics of ligand binding were studied with ITC, the most powerful and sensitive method to elucidate enthalpy and entropy of protein-ligand binding. Theoretical studies were done on the structures to study dynamics and water networks. QM calculations, molecular dynamics simulations and free energy perturbations were the common methods used. To summarize the goal was to understand the subtle changes in binding upon simple changes in the ligand using subatomic X-ray structures and other biophysical methods. The term ligand and compound have been used interchangeably and they mean the same thing.



## 5. Methods

### 5.1. Cloning

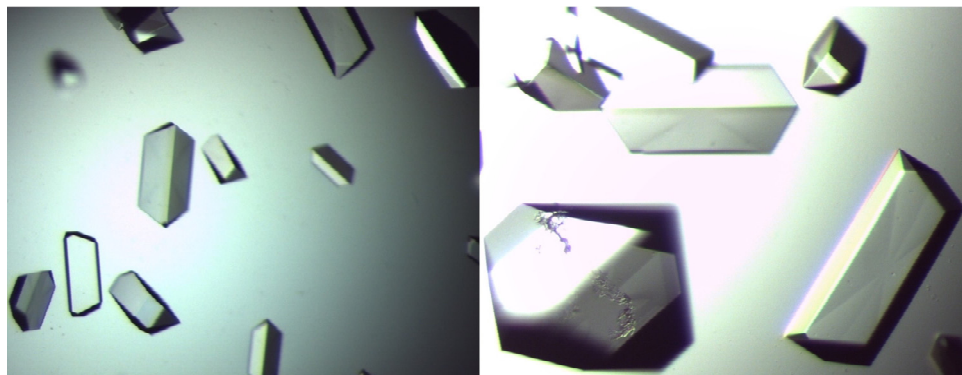
Cloning is basically the process of transferring a gene from one genome /plasmid to another genome/plasmid. The gene of interest is identified, in our case a human galectin-3 CRD gene (LGALS3). Then with the use of restriction enzymes the gene is cleaved from the source DNA and then pasted in the desired plasmid or vector DNA. The recombinant DNA is then transformed into desired expression cells. The galectin-3 CRD wild type gene was previously cloned in pET9a plasmid using NdeI/BamHI restriction sites<sup>117</sup>. Full length galectin-3 is 250 residues; the CRD is residues 113-250. Four mutant gene fragments were ordered from GeneArt, Invitrogen. Mutants Arg144 to Lys/Ser, and Arg186 to Lys/Ser were then sub-cloned into the same pET9a vector using the same restriction sites. Sequencing was done to confirm the positive clones. These positive clones were then transformed into *E. coli* BL21(DE3) expression cells.

### 5.2. Recombinant protein purification

Both the wild type and mutants were expressed and purified the same way. The expression bacterial cells were grown at 37 °C and kanamycin was used for antibiotic selection. Isopropyl thio-galactoside (IPTG) was used to induce the expression and the cells were further grown for four hours and then pelleted. The purification process was same as previously described<sup>21,117,118</sup>. Affinity chromatography was used to purify the protein. A lactosyl-sepharose<sup>118</sup> column was used to purify the proteins from the lysate. Further purification was done by size-exclusion chromatography. Phosphate buffered saline (PBS) was used as the buffer for purification and storing the protein.

## 5.3. Crystallisation

Crystallization of the galectin-3 CRD was first reported in 1998<sup>61</sup>. There has been some optimization to the crystallization condition to achieve very well diffracting crystals as reported in several papers<sup>21,30</sup>. A precipitant solution with 0.1 M Tris pH 7.5, 20% PEG4000, 0.4 M sodium thiocyanate, 5 mM  $\beta$ -mercaptoethanol was used. The hanging drop or sitting drop vapor diffusion methods were used to produce crystals. Crystals appear in 1-2 days and grow to full size in 4-5 days.



**Figure 18:** Well-formed crystals of galectin-3 CRD with lactose

## 5.4. Crystal manipulations

To obtain good data, just getting a crystal is not enough: one needs to manipulate the crystals according to the needs of experiments. While dealing with ligands one needs to soak the crystals in ligand solution. Also, if one needs bigger crystals for room temperature or neutron data collection, seeding methods are used. I will discuss few of the methods I used in my work.

### 5.4.1. Micro-seeding

Micro-seeding is a process where very small seeds are made from a few big crystals (mostly from apo protein crystals). Then these seeds are diluted and used to provide nucleation, which will result in either a few big crystals or several/many small crystals, depending on the dilution. I used this method to get crystals of protein-ligand complexes where spontaneous nucleation was not present because of the DMSO (di-methyl sulfoxide) that the ligands are usually dissolved in. The galectin-3C-lactose crystals were transferred to 50  $\mu$ l of reservoir solution and crushed with

a Seed Bead from Hampton Research. This seed stock was frozen and stored. I used 1:100 dilution from the seed stock to seed new drops.

#### **5.4.2. Macro-seeding and feeding**

This method was used to grow very large crystals for neutron crystallography. The galectin-3C-lactose crystals were washed with the crystallization buffer and then transferred to a larger drop (ideally 20  $\mu$ l or more) with lower precipitant concentration. Then the crystals were monitored, and the drop was fed with fresh protein solution every week. The crystals kept growing, and sometimes nucleation occurred, in which case the crystals were washed and moved to new drops. This worked really well to get large crystals for the neutron data collection.

#### **5.4.3. Soaking**

Ligands were provided in powder form from our synthesis colleagues. They were dissolved in DMSO to make a stock of 50–100 mM depending on the amount. Most of them were insoluble in water so DMSO was necessary, e.g. the compounds in papers I-III. The compounds in papers IV, V, VII and VIII were soluble in water, so they were dissolved in the same buffer as the protein. Soaking is the process of transferring protein crystals to concentrated ligand solution (usually 5-10 mM mixed with reservoir solution) or transferring the ligand solution into the drops having crystals. As a result, ligands can enter the crystals and bind to the proteins. Soaking was usually done for 12-15 hours to get protein-ligand complexes. PEG400 was used to increase solubility for the ligands in papers I and II.

For soaking big crystals for neutron data collection, the crystals were transferred to dialysis buttons (Hampton Research) with protein and reservoir solution, which were covered with a semi-permeable membrane (molecular weight cut-off 10 kDa). The button was then transferred to a well with ligand-reservoir solution (10 mM). This soaking was performed for at least a couple of weeks to fully saturate the proteins with ligands.

### **5.5. X-ray Crystallography**

This was the principal method used for my work. Almost 50 successful cryogenic temperature data sets for unique protein-ligand complexes were collected. Some of them have been used in manuscripts in this thesis. Others need more work in addition to the structural analysis. They have been included in the last chapter.



### 5.5.1. Synchrotrons

The data collection was done at MAX IV Laboratory (MAX II and the new MAX IV), ESRF (European Synchrotron Radiation Facility), Grenoble and DESY (Deutsches Elektronen-Synchrotron) in Hamburg. Beamline I911 at MAX II was used mostly for data collection of the structures in papers I and II. The data collection time was usually 60 minutes with a CCD-detector. The beamlines used at the ESRF were ID23 and ID30, and at DESY the EMBL beamlines P13 and P14 were used. Papers I and IV have data collected from these sources. The BioMAX beamline at MAX IV was used for data collection for papers III and VI. All these beamlines had pixel-based detectors and the data collection time was a few seconds, because of advanced detectors and shutterless continuous rotation method<sup>47</sup>.

### 5.5.2. Data collection

Strategies for data collection are experiment and beamline-dependent. One has to consider several factors for collecting a good and complete data set. Soaked crystals generally diffracted poorly during most of my data collection. Thus, one may have to test multiple crystals to get a good dataset. Technical factors to consider are exposure time, transmission, number of images and wavelength/energy of the beam. Exposure times generally used at DESY and ESRF were 0.02 seconds, whereas at MAX IV it was 0.008 seconds or higher. The combination of exposure time and radiation intensity is important to optimize; higher values produce radiation damage, so one has to choose the optimal combination. Transmission was selected according to suggestions made by characterisation; higher transmission gives intense spots but can cause radiation damage. Thus, optimal values of exposure time and transmission are necessary to get damage free high-resolution data. Choosing the right energy/wavelength is also necessary and depends on how the crystals diffract. Galectin-3 CRD crystals generally diffract very well (around 1 Å). So, choosing a high-energy beam is necessary to go lower in wavelength. In paper III, ligands were prone to radiation damage because of halogen atoms (Br, I), so multiple datasets were collected with varying exposure times and transmission values to get a dataset without radiation damage. The number of images to collect depends on the space group and experiment (anomalous data needs higher number of images). To get complete data sets, 0.1° oscillation and 1800-3600 images were collected.

#### *Cryo data collection*

Data collection done at 100 K is called cryo-data collection. The crystals are frozen in liquid nitrogen and then mounted by the sample changer. A jet of liquid nitrogen keeps the crystals at 100 K, thereby reducing the radiation damage. One has to use a cryoprotectant to avoid ice formation, which can produce its own diffraction pattern. In this work PEG400 (20%) was used as cryoprotectant.

### *Room temperature data collection*

Room temperature data is somewhat tricky to collect, as the chances of radiation damage are high. Thus, one has to adopt several strategies to minimize the damage, e.g. choosing a bigger crystal so that dose of radiation can be spread over a larger area by using helical data collection strategy. However, using a large, defocussed beam is the best way to achieve low damage. Dehydration could be a problem as well, so the crystals are mounted in a loop, which is sealed with the help of a plastic cap with reservoir solution at one end. The room temperature kit from MiTeGen (MicroRT<sup>TM</sup>) was used for this purpose.

### **5.5.3. Data Processing**

Diffraction data processing was done with XDS<sup>119</sup>. Since the space group of galectin-3 CRD crystals are known (P2<sub>1</sub>2<sub>1</sub>2<sub>1</sub>, space group number 19), the input file was modified to have those values before the start of processing. The output CORRECT.LP file was analysed to review several statistical parameters. Most important ones are CC<sub>1/2</sub>, I/σ (signal to noise ratio), completeness of data, and Rmeas<sup>120</sup>. CC<sub>1/2</sub> is the most important statistic values to consider. CC<sub>1/2</sub> is a special case of Pearson correlation coefficient (CC); rather than determining the correlation between two independent datasets, a single dataset is randomly divided into two subsets and CC is calculated from these. CC<sub>1/2</sub> is mathematically explained by following equation:

$$CC_{1/2} = \frac{\sum_{i=1}^n (x - \langle x \rangle)(y - \langle y \rangle)}{\sqrt{\sum_{i=1}^n (x - \langle x \rangle)^2} \sqrt{\sum_{i=1}^n (y - \langle y \rangle)^2}}$$

Where x and y are random subsets of a complete data. CC is correlated to CC<sub>1/2</sub> by following equation:

$$CC = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

RESOLUTION LIMIT	NUMBER OF REFLECTIONS OBSERVED	UNIQUE	POSSIBLE	COMPLETENESS OF DATA	R-FACTOR observed	R-FACTOR expected	COMPARED	I/SIGMA	R-meas	CC(1/2)
4.76	6788	798	802	99.5%	7.3%	6.8%	6784	24.59	7.8%	99.7*
3.38	12443	1334	1334	100.0%	8.2%	7.6%	12442	23.58	8.7%	99.7*
2.76	15679	1703	1704	99.9%	13.2%	12.3%	15674	14.09	14.0%	99.4*
2.39	17984	1991	1991	100.0%	22.5%	19.7%	17982	8.42	23.9%	97.9*
2.14	20382	2223	2224	100.0%	30.1%	25.5%	20382	6.29	31.9%	96.2*
1.95	22406	2447	2447	100.0%	39.9%	34.0%	22404	4.13	42.4%	94.0*
1.81	22860	2659	2661	99.9%	55.9%	50.5%	22855	2.22	59.4%	86.5*
1.69	23812	2842	2842	100.0%	76.2%	77.7%	23805	1.06	81.0%	69.1*
1.60	23687	3005	3017	99.6%	94.9%	106.2%	23665	0.50	101.1%	49.1*
total	166041	19002	19022	99.9%	16.4%	15.1%	165993	6.65	17.4%	99.7*

**Figure 19:** Snapshot of data processing statistics from CORRECT.LP, showing important parameters to consider.

Completeness of data should be as high as possible; more than 90% completeness is desired.  $I/\sigma$  should be ideally higher than 1 for the highest resolution shell but one has to consider  $CC_{1/2}$  value as well. So, the resolution cut-off is based on combination of  $CC_{1/2}$  and  $I/\sigma$ . CC values should be above 0.3<sup>49</sup> for the data to be relevant. As one can see here, the resolution cut-off was at 0.5 for  $I/\sigma$  since the  $CC_{1/2}$  values were 99%, which means the data are still relevant. The processed data were scaled using Aimless from the CCP4 suite<sup>121</sup>.

#### 5.5.4. Refinement and model building

Refinement and model building were done using the Phenix suite<sup>122,123</sup> and COOT<sup>124</sup>. Molecular replacement or Fourier synthesis were used for solving the structure. The lactose-galectin-3 CRD structure was used as the template and water, ligand and hydrogens were stripped. Fourier synthesis was performed by doing rigid body refinement at lower resolution then gradually increasing the resolution and finally switching to real space refinement. Ligands were built with eLBOW in Phenix<sup>125</sup> or Acedrug<sup>126</sup> in CCP4 by using SMILES strings or drawing a 2D structure in COOT ligand builder. The structures were further refined until the R-factors converged. Individual anisotropic B-factors option was selected (except for hydrogens), as the data were of higher resolution than 1.5 Å. The B-factor, atomic-displacement parameter or Debye-Waller factor quantitates the uncertainty for each atom<sup>127–129</sup>. The higher the B-factor, the higher is the mobility and hence uncertainty. It is an important parameter in analysing the structure and its dynamics. Highly flexible parts have higher B-factors and vice-versa. The B-factor also gives information about errors in model building<sup>129</sup>. The B-factor is given by following equation:

$$B_i = 8\pi^2 U_i^2$$

Where  $U_i^2$  is mean square displacement of atom i.

Occupancy of ligands is important, and complete occupancy of the ligand is desirable. If the occupancy is very low, it is important to recollect data with longer soaking times. I had to re-collect several datasets to get complete occupancy.

Model building was done in COOT, each residue was checked for geometry and electron density. The final structures were checked for quality by PDB\_REDO<sup>130</sup> before deposition. 24 datasets have been deposited in the Protein Data Bank so far, with many more remaining.

## 5.6. Neutron crystallography

The major limitation of X-ray crystallography is inability to see hydrogens in most cases. This is where neutron crystallography comes into play. X-rays interact with electrons and give electron density maps. Therefore hydrogens (H) with one electron or protons ( $H^+$ ) with none have poor or no scattering power and will be almost impossible to see, except at very high resolution X-ray structures (e.g. PDB id: 5D8V), where still all the hydrogens are not visible<sup>131,132</sup>. Neutrons on the other hand interact with nuclei giving nuclear scattering length density maps. Particularly when  $^1H$  are replaced by  $^2H$ , also known as deuterium (D), the higher neutron scattering power of D makes it easier to visualize<sup>131,133–137</sup>. Coherent scattering from D is similar to C and O.  $^1H$  has the additional disadvantages of a high incoherent scattering cross-section – giving rise to noise – and a negative scattering length leading to negative peaks in the maps.<sup>133</sup>

X-rays interact with electrons and give electron density maps. Therefore hydrogens with one electron (H) or protons ( $H^+$ ) with none have poor or no scattering power and will be almost impossible to see<sup>131</sup>. Neutrons, on the other hand, interact with nuclei, giving nuclear scattering length density maps. Particularly when  $^1H$  are replaced by  $^2H$ , also known as deuterium, (D) in the proteins the higher scattering power of D makes it easier to visualize<sup>131,133–137</sup>. The coherent scattering from D is similar to C and O.  $^1H$  has the additional disadvantages of a high incoherent scattering cross-section – giving rise to noise – and a negative scattering length, leading to negative peaks in the maps<sup>133,138</sup>. Data from neutron diffraction helps us see hydrogen bonds, protonation states and directionality of the bonds as well<sup>133</sup>. This is a great complementary method to X-ray crystallography, and given how important H-bonding, water networks and protonation states are in ligand binding, catalysis *etc.*, the importance of this method is even greater. We have utilized this technique to answer certain questions involving key H-bonds between solvent and the ligands, while also looking at the direction of the bonding.

### 5.6.1. Deuterated protein production and crystallization

For neutron crystallography, perdeuterated protein is very helpful. The labelling is achieved by growing cells that overexpress the protein using heavy water D<sub>2</sub>O instead of H<sub>2</sub>O and deuterated d7-glucose or d8-glycerol is used as an energy source. In our work, the cells were gradually adapted to increasing concentrations of D<sub>2</sub>O using M9 minimal media. Expression of the protein was carried using IPTG prepared in D<sub>2</sub>O. The cells were pelleted and then the purification was carried out as described previously using non-deuterated buffers. The purified protein was then exchanged to deuterated buffer and stored at -80 °C. The detailed protocol of perdeuteration and crystallization of galectin-3CRD is explained in paper VII. The macro-seeding method was used to produce big crystals, as described in section 5.4.2. Lactose was added to protein as it helped with crystallization.

### 5.6.2. Data collection and data processing

Neutron sources have a low brilliance, so the data collection times are longer (a few days or weeks), but fortunately there is no radiation damage. One needs large crystals to measure diffraction patterns because the diffracted beam intensity is directly proportional to volume of crystal and incident beam intensity<sup>139,140</sup>. The reason for the less intense neutron beam is low flux (number of particles (neutrons or photons) cm<sup>-2</sup>s<sup>-1</sup>) compared to X-rays. For comparison, neutron sources have a flux of 10<sup>6</sup> to 10<sup>8</sup> while X-rays at synchrotron sources have a flux of 10<sup>16</sup>.<sup>138</sup> This is the primary reason for high data collection times. There are very few neutron instruments available, and one data collection can take up to two weeks, so the planning of experiments is crucial. We collected data at LADI (Institut Laue-Langevin, Grenoble)<sup>141</sup>, BioDiff<sup>142</sup> (Heinz Maier-Leibnitz-Zentrum, Munich) and MaNDi<sup>143</sup> (Oak Ridge National Laboratory, Oak Ridge, TN). The crystals are mounted in appropriately-sized quartz capillaries and some reservoir solution is added at both ends to avoid dehydration. The capillaries are sealed at both ends by using wax. The exposure time can be anywhere between 30-90 minutes at BioDiff, and a few hours (up to 24 hours) at LADI and MaNDi. At LADI or MaNDi the number of images collected is much smaller compared to X-ray data collection, as these instruments use the quasi-Laue method<sup>144</sup>. BioDiff is the only monochromatic source of neutron among the mentioned sources. For example, at LADI one typically collects images by rotating 7° between images.

Data processing involves similar steps to X-ray crystallography, like indexing of spots, integration and scaling of intensities. Lauegen and LSCALE in the Daresbury Laue Suite<sup>145</sup> are specialized software for that purpose. Typically, we were provided with scaled data files by the beamline scientists, which is common in the field of neutron crystallography. Data processing involves similar steps to X-ray

crystallography, like indexing of spots, integration and scaling of intensities. Lauegen and LSCALE in the Daresbury Laue Suite<sup>145</sup> are specialized software for that purpose.

### 5.6.3. Joint X-ray/neutron refinement

Once the neutron data are collected, the same crystal is used to collect X-ray data, as there is no radiation damage from neutron beam. The X-ray data are then processed and the model is refined against X-ray data first (at similar resolution to neutron data). Neutron data were of slightly lower resolution (1.7 Å or lower). Then the neutron data file is added to the refinement and the model is jointly refined against X-ray and neutron data. Phenix was used for refinement of models<sup>146–148</sup>. Deuterium atoms were added to the model including the solvent and all the deuteriums were chosen to be individually refined with restraints<sup>149</sup>. Deuterium was added to the model, including the solvent and ligand. Ligands were not deuterated so aliphatic carbons were manually modified to delete deuteriums. Waters that were clearly visible in the neutron map were added. After few refinement cycles, the rest of the high-resolution X-ray data was added, and the model was refined with anisotropic ADPs.

## 5.7. Other methods used in papers

Apart from the crystallography methods, several other methods were used to complete the study, like fluorescence polarization (FP), isothermal titration calorimetry (ITC), NMR studies, and theoretical calculations (Quantum Mechanical (QM), Molecular Mechanics (MM) and Grid Inhomogeneous Solvation Theory (GIST) calculations). These experiments were mostly performed by collaborators. FP was used to find the binding affinity of all the synthesized compounds. ITC was measured on selected compounds to illustrate their binding thermodynamics. NMR was used to investigate the protein-ligand dynamics. Theoretical studies were performed with the crystal structures to study dynamics, entropy and role of water in binding. All these data were analysed together to give a complete picture of protein-ligand binding.



## 6. Results and Discussion

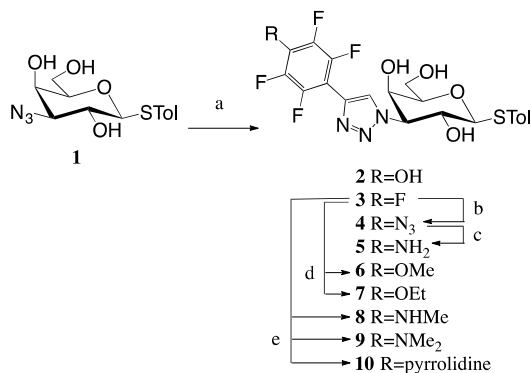
In the following sections, I have summarized my results paper wise. These discussions mostly cover aspects of paper that involves my work. Papers I and II involve FP data, X-ray structures and theoretical calculations to show the effects of substituting at single positions in the ligand. Papers III and IV include FP data, X-ray structures, ITC and theoretical calculations. Papers V and VI include, FP data, X-ray structures, ITC and biological data. Papers VIII and IX include FP data and X-ray structures only; the ITC data is incomplete so not included here. Paper VII includes expression, purification and crystallization of perdeuterated galectin-3 CRD.

In the final section 6.10, I have included neutron data for two series of compounds in this section and not as manuscript, as the data is either negative or incomplete. I have explained the results obtained by neutron diffraction and role of H-bonding between ligand and protein. For the second series of compounds, only one neutron structure is available, hence the story is incomplete.



## 6.1. Paper I

**Substituted polyfluoroaryl interactions with an arginine side chain in galectin-3 are governed by steric-, desolvation and electronic conjugation effects**

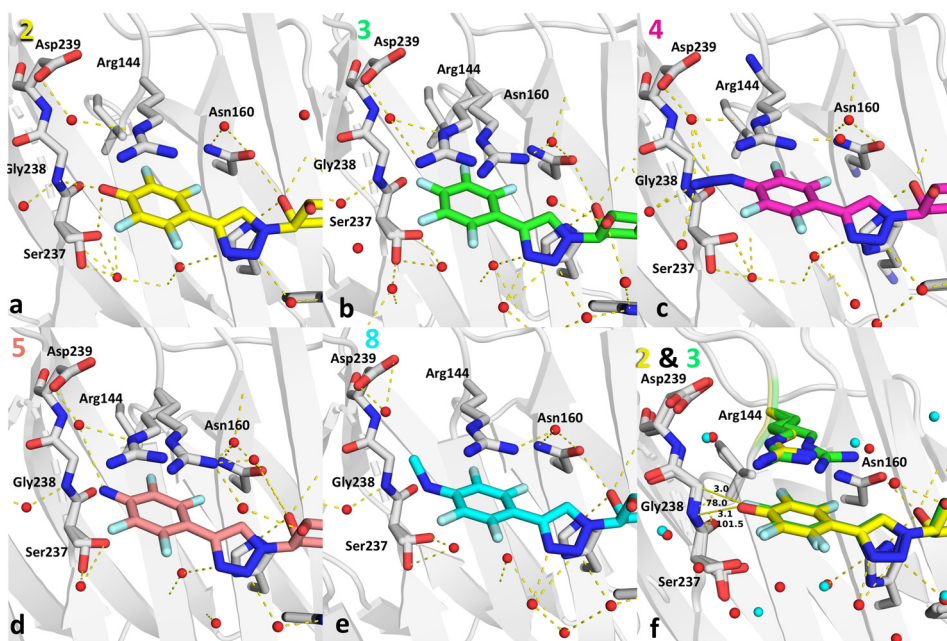


**Figure 20:** Schematic view of the compounds used in this work

In this paper we have explored the binding subsite A (for nomenclature, refer to Section 3.1) close to Arg144.

Nine 3-(4-(2,3,5,6-tetra- fluorophenyl)-1,2,3-triazol-1-yl)-thio-galactosides with different para substituents were synthesized, as shown in the image on the left. Their affinities were determined with FP. X-ray structures for five of them were solved and analysed.

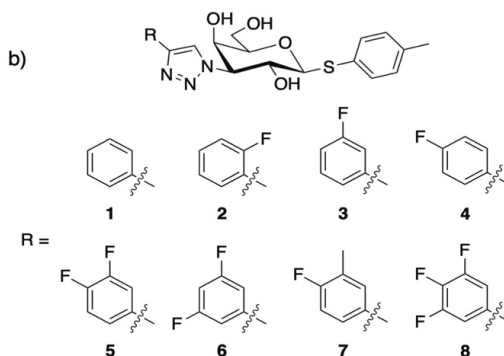
Binding affinities were explained using the structures and quantum mechanical (QM) calculations. The larger substituents at para position had poorer affinity because they were too large for the binding pocket near Arg144. Compound 3, with fluorine at the para position, had the highest affinity because fluorines interact with the backbone of Ser237-Gly238. For other compounds the affinity is governed by the desolvation penalty, which disfavors polar substituents, and cation- $\pi$  interactions between Arg144 and the fluorophenyl triazole group.



**Figure 21:** Structures of compounds 2-5 and 8, showing key residues and polar contacts

## 6.2. Paper II

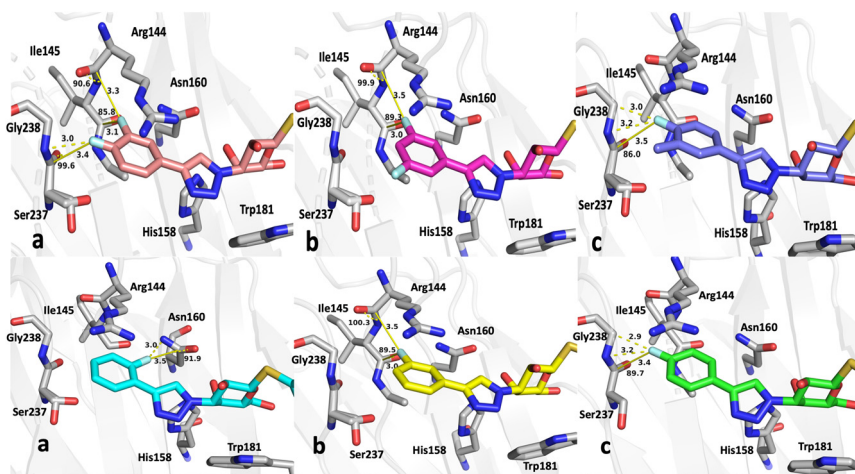
### Structure and energetics of ligand–fluorine interactions with galectin-3 backbone and side-chain amides – insight into solvation effects and multipolar interactions



**Figure 22:** Schematic view of the compounds used in this work

In this paper, we explored multipolar fluorine-amide interactions with protein backbone and side chain amides. These interactions are thought to play an important role in the potency of protein-ligand interactions. The fluorine position was varied in the phenyltriazole group of the ligands. The affinity data showed fluorine at the meta (**3**) and para (**4**) positions enhanced affinity, while ortho fluorine had poorer affinity. Having fluorines at two positions in **5** and **8** further enhanced the affinity. We solved eight high-resolution X-ray structures to elucidate the interactions. The structures showed fluorines forming orthogonal multipolar interactions with nearby backbone amides and side chain amides (in **2**). Fluorine-backbone amide interactions were stronger compared to fluorine-sidechain amide.

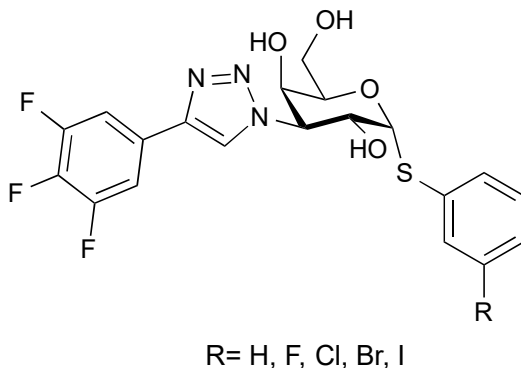
This was further confirmed by quantum mechanics calculations. However, these calculations also showed that the affinity enhancement is not primarily associated with the predicted fluorine-amide interactions, but that desolvation and dispersion effects play a larger role in case of multi-fluorine containing ligands.



**Figure 23:** Structures of the compounds **2** (cyan), **3** (yellow), **4** (green), **5** (brown), **6** (magenta) and **7** (purple), showing fluorine-amide interactions.

### 6.3. Paper III

#### Structural and thermodynamic studies on halogen-bond interactions in ligand-galactin-3 complexes: electrostatics, solvation and entropy effects

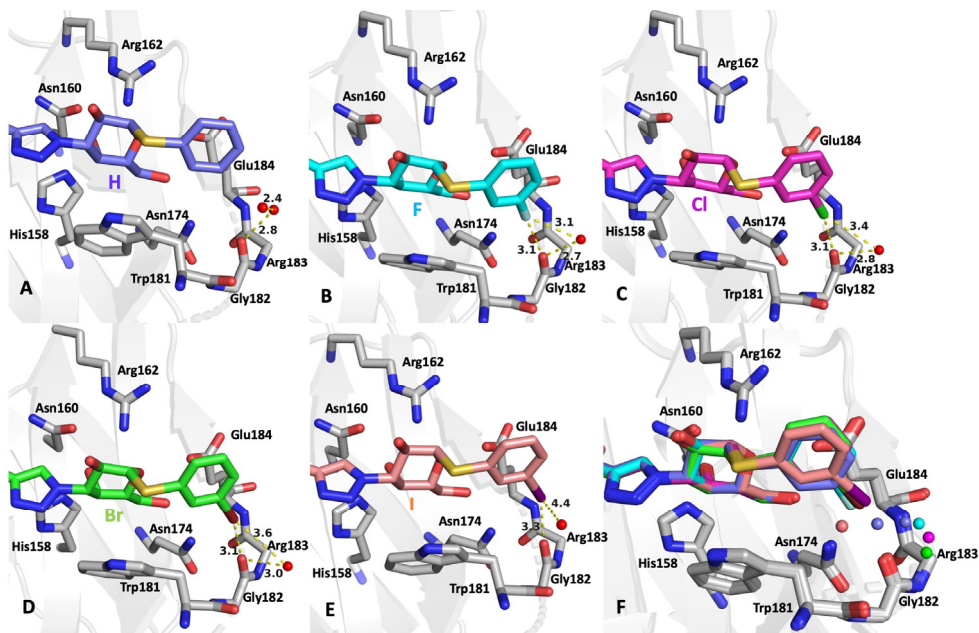


**Figure 24:** Schematic view of the compounds used in this work

In this work, we have explored the structure and energetics of halogen-bond interactions in protein-ligand binding. Halogens, owing to their anisotropic electron distribution, have a so-called  $\sigma$ -hole that is positively charged. Thus, halogens can act as electron acceptors and interact with electron donor groups like carbonyl oxygen, with interaction distances that are shorter than the sum of the van der Waals radii of the atoms. Five compounds were synthesized, with H, F, Cl, Br and I at the position R shown in the image. The idea for synthesizing these compounds came from recent studies that identified a novel halogen bond to the carbonyl oxygen of Gly182 in the galectin-3 CRD<sup>150</sup>. The affinity as determined by FP showed an increase in affinity with increasing size of halogen. However, fluorine does not have a  $\sigma$ -hole and hence does not show halogen bonding capabilities because of its small size and high electronegativity.

X-ray structures of all five compounds were obtained in complex with the galectin-3 CRD. Structural analysis showed the halogen atoms forming halogen bonds with the carbonyl oxygen of Gly182 in all the structures. The distance between the halogen atoms and the carbonyl oxygen remained the same, and the phenyl group moved slightly to compensate for the increasing size of the halogen atom. The distances, as expected, were shorter than the sum of their van der Waals radii at 3.1 Å. Structural analyses also showed two water molecules present in the complex with the unsubstituted compound that moved away gradually in the other structures, with one water completely displaced by all the halogens, while the second water molecule was displaced gradually as a function of the halogen size. In I the water

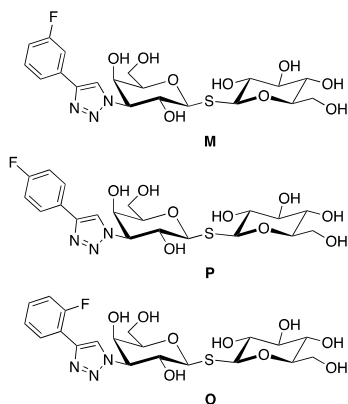
molecules is displaced completely. Halogen atoms are also able to bridge water molecule which is simultaneously H-bonded to a residue. Here Cl, Br and I formed such interaction by bridging a water molecule to carbonyl group of Gly182.



**Figure 25:** Structure of the ligand-galectin-3 CRD complexes showing the halogen-carbonyl oxygen distance and the water molecules. Panel F shows a superimposed view, and the water molecules are coloured the same as their respective ligands.

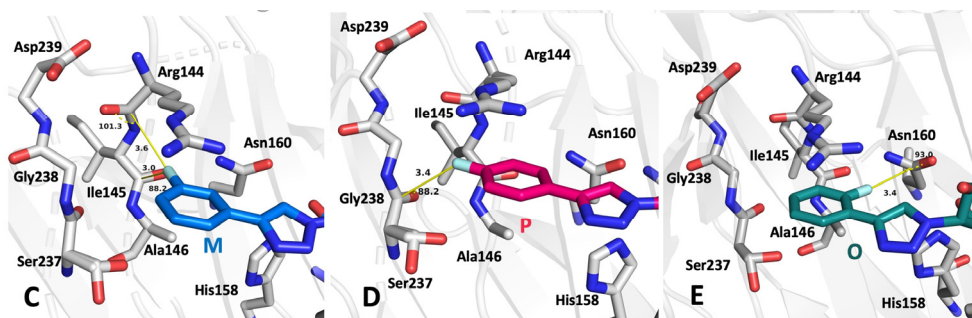
## 6.4. Paper IV

### Entropy–entropy compensation between the conformational and solvent degrees of freedom fine-tunes affinity in ligand binding to the galectin-3 CRD

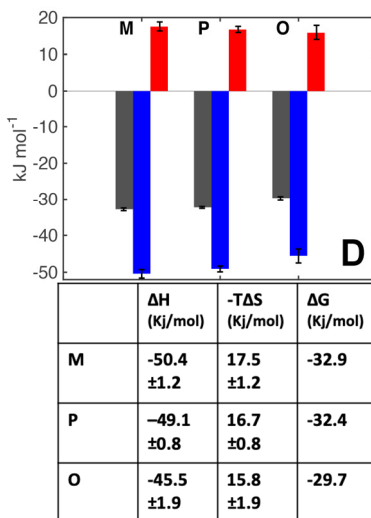


**Figure 26:** Schematic view of the compounds used in this work

This paper is continuation of Paper II. Here we synthesized soluble versions of the ortho(O), meta(M) and para(P) fluorinated compounds from Paper II by adding a glucose, which helped to perform ITC on these compounds. X-ray structures were solved for all three ligand-protein complexes. The binding of the fluorophenyl triazole group is identical to that of compounds 2, 3 and 4 from paper II. They bind near Arg144 and the fluorine atoms form very similar multipolar interactions with backbone or side chain amides as in the monosaccharides. Our main goal was to understand the thermodynamics associated with these interactions, for which NMR and ITC was performed.



**Figure 27:** Image showing a closeup view of M(blue), P(red) and O(green) near Arg144



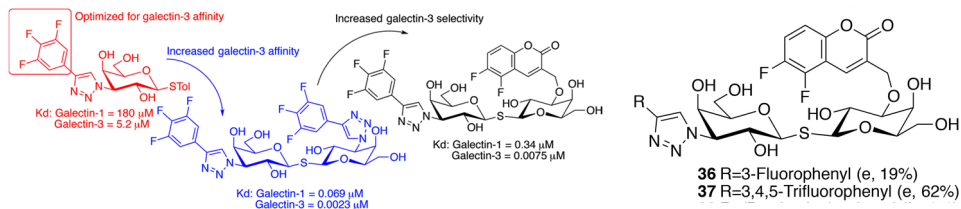
**Figure 28:** ITC data for the three compounds. Entropy (red), enthalpy (blue) and free energy change (gray) are plotted. The table below the graph shows the values.

The ITC data show that the binding is mostly enthalpically driven. O shows a slightly lower enthalpic contribution. Entropic contributions are similar, but the data show a trend towards enthalpic-entropic compensation. Overall the ligands show similar thermodynamic signature. However, NMR data and theoretical calculations (refer to paper IV) suggest entropy-entropy compensation.



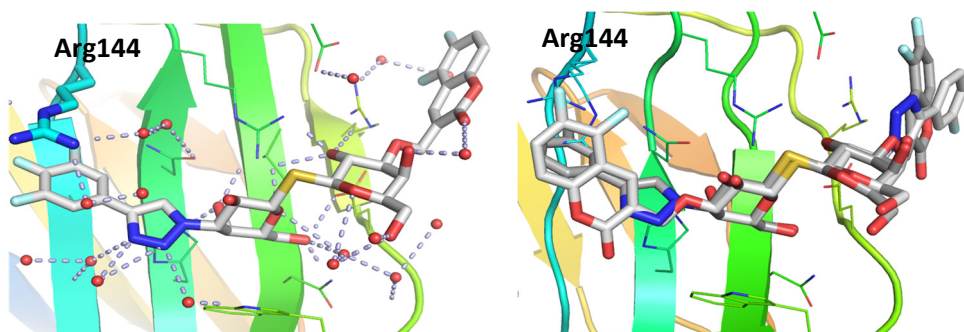
## 6.5. Paper V

### Systematic tuning of fluoro-galectin-3 interactions provides thiodigalactoside derivatives with single-digit nM affinity and high selectivity



**Figure 29:** Schematic view of the compounds used in this work

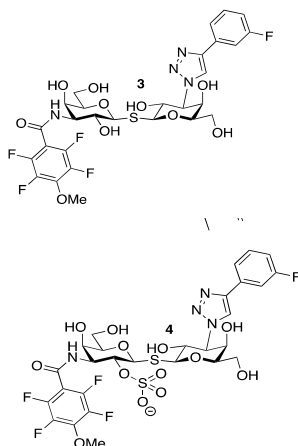
In this work, we showed how changing certain parts of the ligand scaffold provides specificity and/or affinity. As seen in the image above, adding a trifluorophenyl group near subsite A increases affinity for both galectin-3 and galectin-1, while adding a coumaryl group near binding subsite E enhances selectivity for galectin-3 over galectin-1. The two compounds were high affinity binders, **36** with 27 nM and **37** with 4 nM affinity. X-ray structures were solved for the two ligand-galectin-3 CRD complexes. Structures showed that the asymmetric ligand **37** with a trifluorophenyl group at one end always binds in single conformation near Arg144 (Fig. 23), while compound **36** with a monofluorophenyl group binds in two conformations. In one conformation the monofluorophenyl group binds near Arg144, while in second conformation it binds near subsite E. This suggests that having multiple fluorines near Arg144 selects for single binding conformation while also providing increase in affinity. Addition of a coumaryl group enhances selectivity, as shown by FP data, but it does not provide any conformational bias towards a single binding mode.



**Figure 30:** Structures of **37** (left) and **36** (right). **36** shows a single binding conformation because of three fluorines interacting with peptide bonds. Compound **36** is in two conformations because of its single fluorine.

## 6.6. Paper VI

### A non-permeable high-affinity sulfated ligand for selective extra-cellular galectin-3 inhibition



**Figure 31:** Schematic view of the compounds used in this work

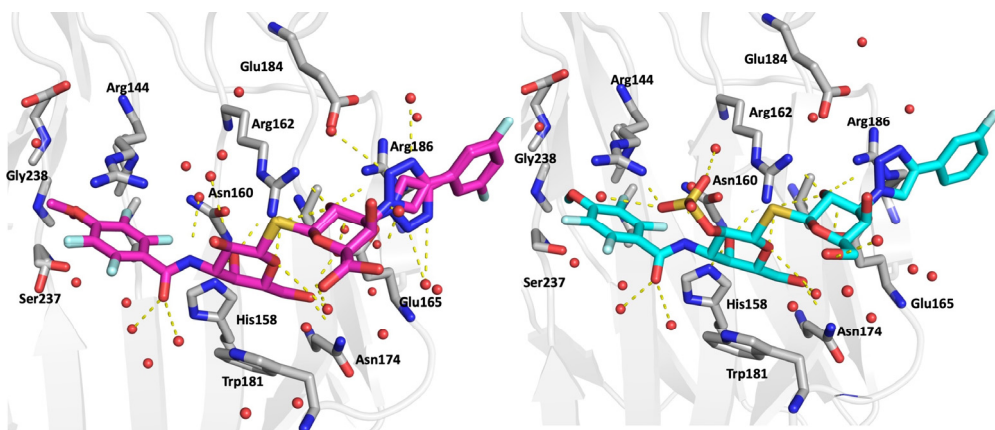
In this paper, we used thiodigalactoside ligands that exhibit nanomolar affinity for galectin-3. Earlier studies<sup>151</sup> showed that the amide linker pulls Arg144 closer to galactose 2'-OH. A sulfate group was introduced to exploit this interaction between Arg144 and the ligand. Compound **3** had affinity of 18 nM and **4** had 6 nM, as calculated from FP data. X-ray structures for the two ligand-protein complexes were solved. Structural analyses showed how the additional sulfate group pulled Arg144

by forming an electrostatic interaction, while also making H-bonds with water molecules. The tetrafluoro-benzamide group bound near Arg144 while the phenyltriazole group bound near Arg186 in subsite E. Compound **3** has dual conformations near Arg186 while **4** has only one, which indicates that the sulfate group restricts the change in binding orientation.

ITC data suggest enthalpic binding with little or no entropic contribution. The  $K_d$  from ITC is higher than FP, but the correlation between **3** and **4** is similar.

	$\Delta H$ (kJ/mol)	$-T\Delta S$ (kJ/mol)	n	$K_d$ (nM)
<b>3</b>	$-38.5 \pm 1.1$	$0.0 \pm 0.7$	1.03	$210 \pm 31$
<b>4</b>	$-39.4 \pm 1.1$	$-1.1 \pm 0.6$	1.04	$95 \pm 17$

The addition of a sulfate group also makes the ligand impermeable to the cell membrane and it can thus be used to target extracellular galectin-3.



**Figure 32:** Image showing binding of **3**(magenta) and **4**(cyan) in the galectin-3 CRD binding pocket. Key residues and polar contacts are shown.

## 6.7. Paper VII

### **Perdeuteration, crystallization, data collection and comparison of five neutron diffraction data sets of complexes of human galectin-3C**

In this paper, we describe expression of deuterated galectin-3 CRD and its crystallisation to produce bigger crystals for neutron diffraction. For expression of deuterated protein, minimal media was used to grow the bacterial cells, which were gradually adopted to increasing concentrations of deuterium oxide (D<sub>2</sub>O). The energy sources used were deuterated d7-glucose or d8-glycerol. Cells were adapted to D<sub>2</sub>O using methods similar to those published previously. M9 minimal medium was used for protein expression. A single colony of *E. coli* BL21(DE3) cells containing the gal3CRD\_pET9a plasmid was grown overnight on M9 agar plate. This was used to inoculate 50 ml of 20% D<sub>2</sub>O M9 medium (with nondeuterated glycerol/glucose) to an OD600 of 0.1, which was then grown for 24 h. The 20% D<sub>2</sub>O culture was used to inoculate 50 ml 100% D<sub>2</sub>O M9 medium (with nondeuterated glycerol/glucose) to an OD600 of 0.1, and the culture was grown for 24 h.

The 100% D<sub>2</sub>O culture was used to inoculate 200 ml 100% D<sub>2</sub>O M9 medium with glycerol-d8/glucose-d7 to an OD600 of 0.1. To avoid transfer of medium without glycerol-d8/glucose-d7, the cells needed for inoculation were pelleted and the medium was discarded. The cell pellet was then used for inoculation and the culture was grown overnight.

The 200 ml 100% D<sub>2</sub>O /glycerol-d8 culture was used to inoculate 2\*1 l of 100% D<sub>2</sub>O +glycerol-d8/glucose-d7 M9 medium to an OD600 of 0.1. At an OD600 of 0.5, IPTG (prepared in D<sub>2</sub>O) was added to a final concentration of 0.5 mM and induction was continued for 12 h. Cells were harvested at 8000g for 20 min at 20°C. Each pellet (from 1 l culture) was resuspended in 10 ml MEPBS and stored at -80°C.

The purified protein was used to produce larger crystals for neutron diffraction. Dgal3CRD-lactose protein was crystallised, and the crystals were transferred to new drop with low PEG concentration to avoid new nucleation. This drop was then provided with 3-5 µl of protein every week, which lead to growth of existing crystal.

Some of these crystals were used to soak with compounds 3 and 4 from paper VI, and neutron diffraction data were collected. Refinement and results for these data is discussed in section 6.10.

## 6.8. Paper VIII

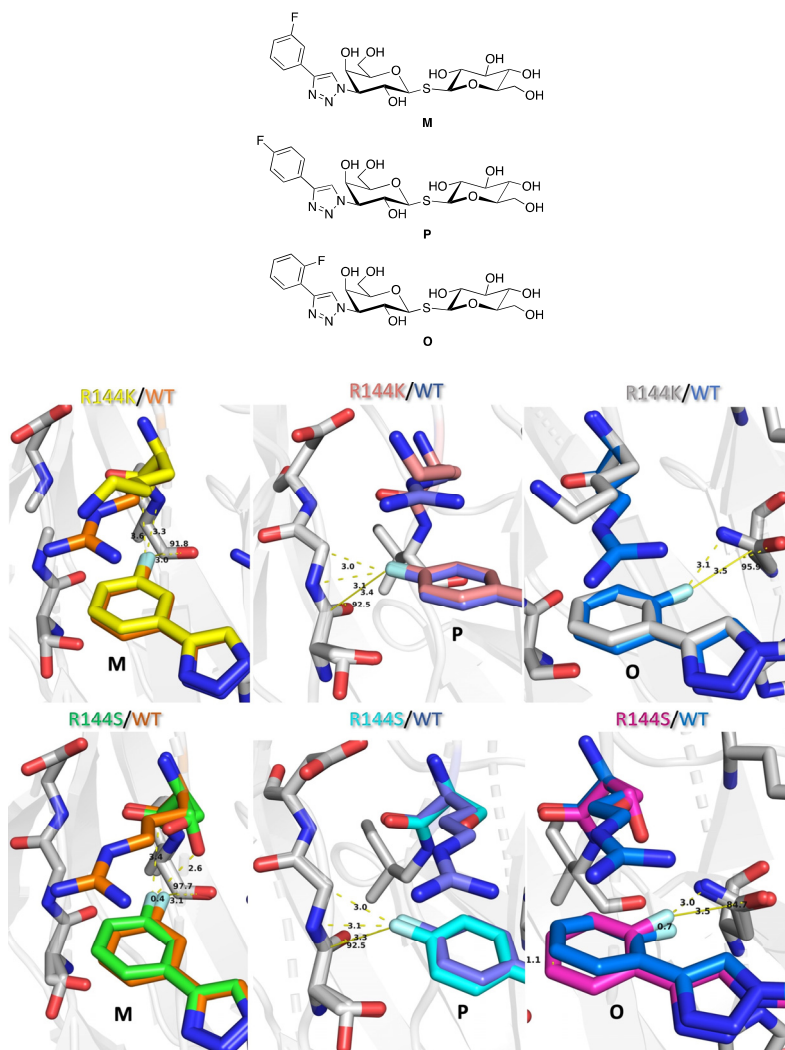
### Structural perspective of Arg144 mutants of galectin-3 CRD in ligand binding: Role of Cation- $\pi$ interactions

Ligand	M	P	O
Protein	Affnity( $\mu$ M)	Affnity( $\mu$ M)	Affnity( $\mu$ M)
Gal3WT	0.46 $\pm 0.06$	0.6 $\pm 0.05$	1.9 $\pm 0.09$
Gal3R144K	2.05 $\pm 0.1$	1.67 $\pm 0.25$	9.02 $\pm 1.58$
Gal3R144S	7.0 $\pm 0.23$	2.3 $\pm 0.13$	10.3 $\pm 1.31$

This paper is continuation of Paper IV, where we showed binding thermodynamics of three ligands O, M and P to wildtype galectin-3 CRD. In this paper we have shown the structural aspects of binding of the same ligands to Arg144 mutants of galectin3CRD. Arg144 was mutated to either Lys (R144K) or Ser (R144S). FP data and X-ray structures were obtained for the mutant-ligand complexes. The binding affinity and structural aspects were compared to wild-type galectin-3 CRD-ligand complexes.

The binding affinity for the compounds decreased from wild-type to mutants, with R144S mutant having lowest affinity for the ligands. The decrease in affinity was not drastic but it was significant. This was expected, as Lys is chemically similar to Arg so the effect is less pronounced for R144K mutants. The results suggest that cation- $\pi$  interactions between Arg144 and the fluorophenyl group are important for binding.

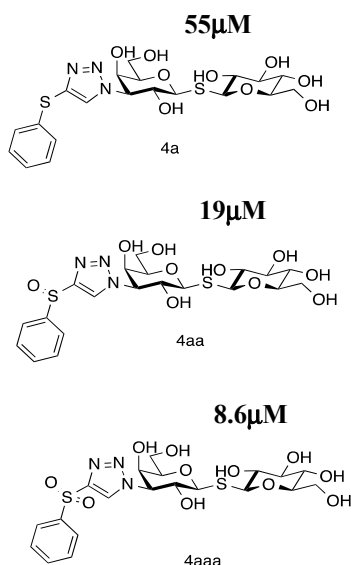
For wild-type, M had the highest affinity but for mutants P has the highest affinity. Also, M shows the most change in affinity between wild-type and mutants, up to a 14-fold difference between wild-type and R144S. This suggests that cation- $\pi$  interactions affect the affinity for M more than for other compounds.



**Figure 33:** Top panel shows Chemical structure of the ligands used for the study. Bottom panel showing closeup images showing binding of the fluorophenyl group near Arg144 in the binding pocket. The structures for wild-type are compared with mutants. As can be seen, binding of Lys mutant is identical to wild-type, only in Ser mutant there are subtle changes to binding.

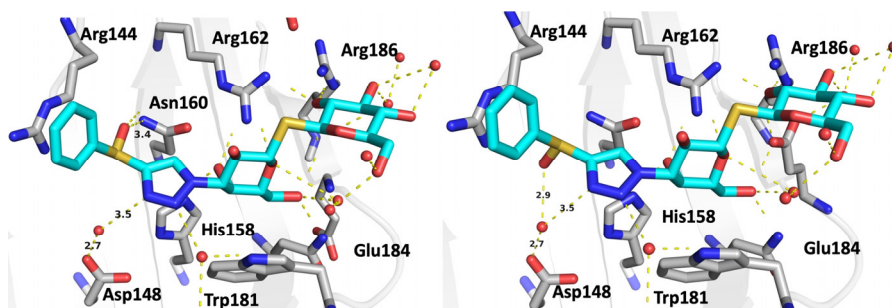
## 6.9. Paper IX

### Ligand sulfur oxidation states stepwise alter ligand-galactin-3 complex conformations



**Figure 34:** Compounds chemical structure and their affinity is shown.

In this paper, three variants of sulfide compounds were synthesized by stepwise oxidation to study the interaction of sulfur oxidation state on the binding affinity and thermodynamics in complex with galectin-3 CRD. We expected the phenylsulfide group to bind near subsite A so that the phenyl group would extend towards Arg144. The affinities of the compounds were determined by FP (shown on left) and the affinity increased with the oxidation state of the sulfur atom. Compound **4aaa** (sulfone) with two oxygens attached to the sulfur atom has the highest affinity. Sulfoxide moiety in compound **4aa** has a chiral centre, so it is supposed to bind in two configuration. X-ray structures were solved to analyse the binding of these ligands. The structural analyses show the sulfide group binding near the Arg144 with the phenyl group pointing towards solution.

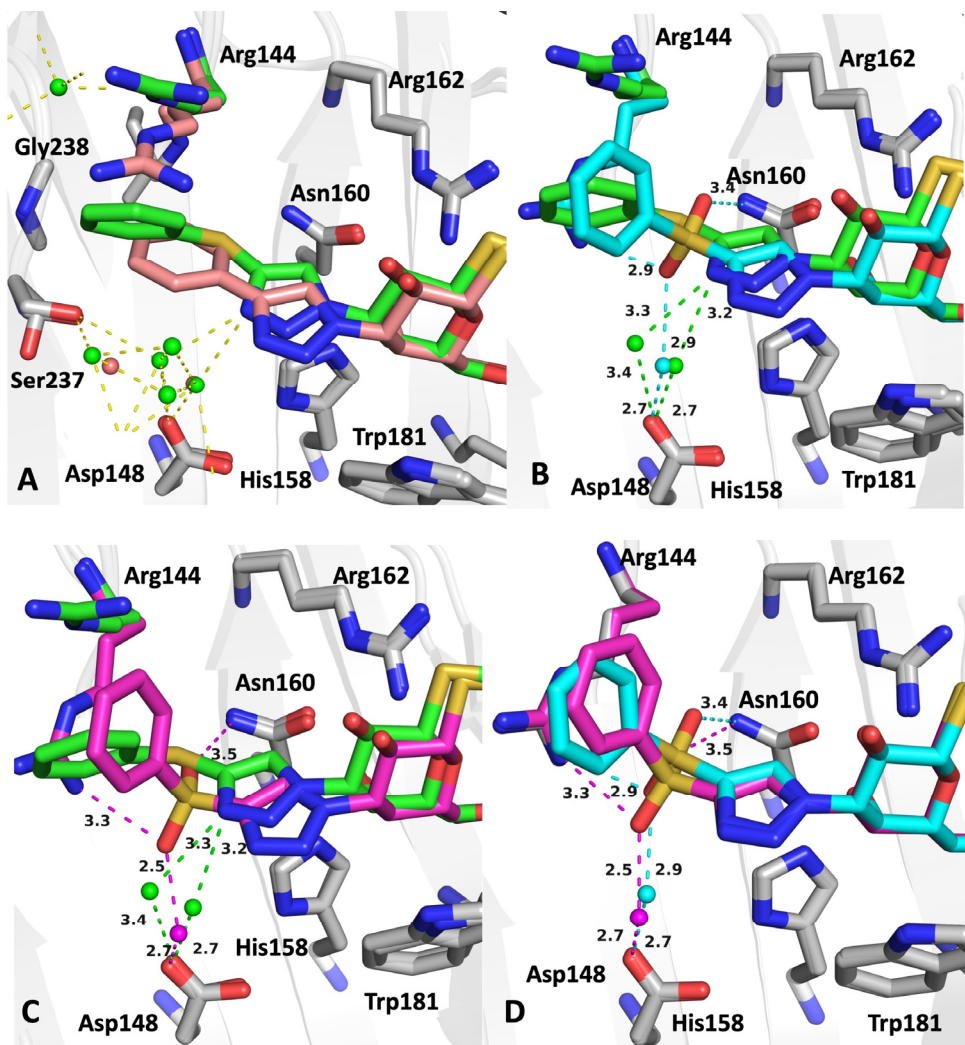


**Figure 35:** Binding of two enantiomeric configuration in **4aa**

The Arg144 maintains cation- $\pi$  interactions with the phenyl group even when the phenyl group in compounds **4aa** and **4aaa** (sulfone) change binding conformation as shown in images below. Oxidation of sulfur makes the phenyl group point orthogonally to the sulfoxide or sulfone; this results in shift of Arg144 such that it still forms cation- $\pi$  interactions. Binding of **4aa** is similar to **4aaa** because of the two enantiomeric configurations.

The oxygen in **4aa** forms H-bond with a water that is also bound to Asp148 and the oxygen in second enantiomer points towards Asn160. In the **4aaa** structure, one oxygen has similar interaction to **4aa** while the other oxygen interacts with Arg144 and Asn160. The structural analysis show that that oxidation of the sulfur group leads to more interactions with solvent and other residues.





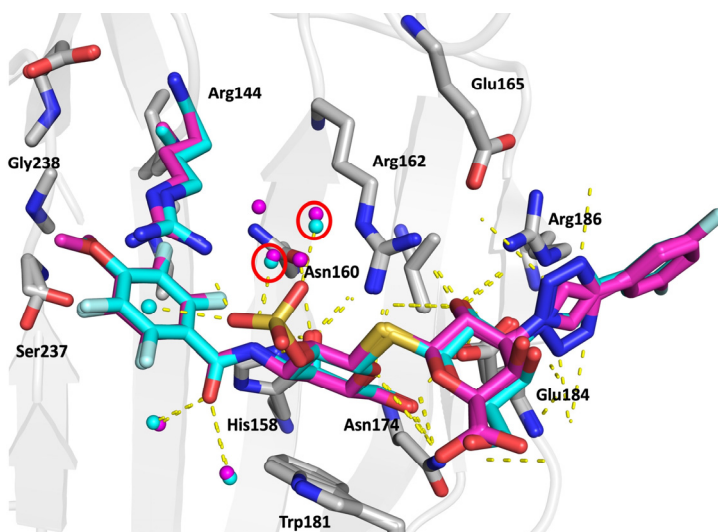
**Figure 36:** Closeup view of the binding of the compounds near Arg144 in the binding pocket. A) Comparison of 4a and unsubstituted phenyltriazole B) 4a (green) and 4aa (cyan)(both configurations together in the pocket. C) 4a (green) and 4aaa (magenta) superimposed. D) 4aa (cyan) and 4aaa (magenta) superimposed. Key polar contacts are shown and colour-coded green for 4a, cyan for 4aa and magenta for 4aaa

## 6.10. Results not yet included in manuscripts

### 6.10.1. Neutron data

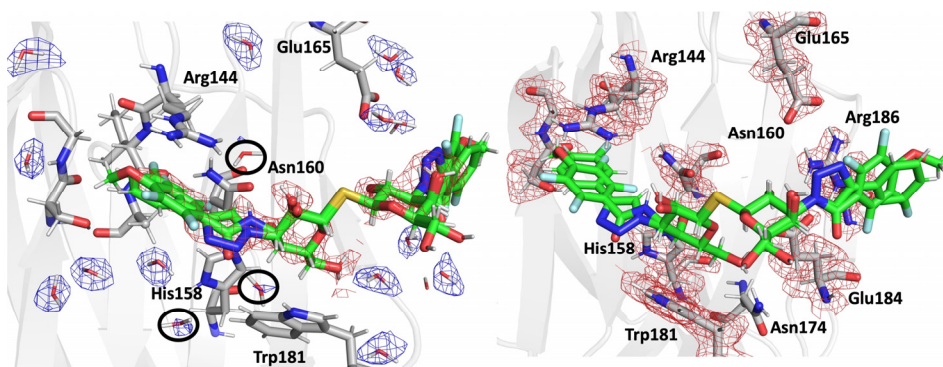
#### *Elucidation of H-bonding and role of waters in Paper VI compounds*

Neutron data were collected for compounds **3** and **4** from Paper VI. We wanted to explore some key water molecules involved in binding, their interactions and orientations of H-bonds. In Figure 28 below, key water molecules (in red circles) are coloured similarly to the ligand complex they are bound to. As can be seen from the image, the water molecule in **3** (magenta) close to the ligand is displaced in **4** (cyan) because of the sulfate groups.



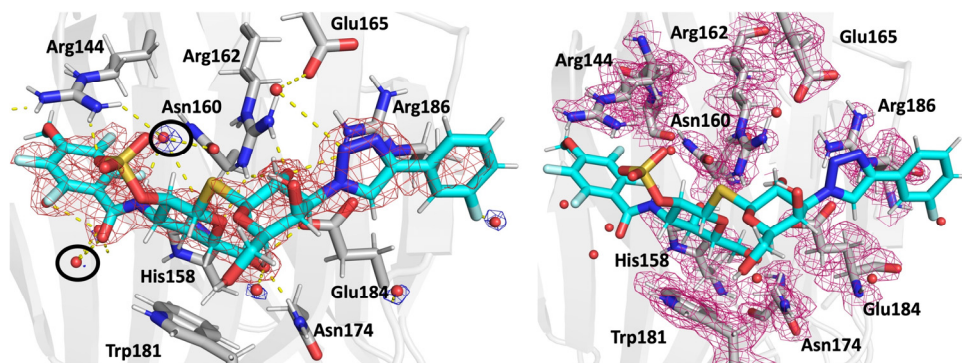
**Figure 37:** Cryogenic x-ray structures of **3** (magenta) and **4** (cyan) compared to show key water molecules involved. Water molecules are also colour coded, magenta in **3** and cyan in **4** to provide the comparison.

Sulfate group makes another polar contact with a water molecule and Arg144 (coloured cyan). We wanted to see the direction of the H-bonding these water molecules are involved in and if there is any change in the two ligand complexes. The data collection for compound **3** is described in Paper VII. The data was processed and refined as described previously in the Methods section. The protein nuclear density map was very good for both the complexes, but ligand nuclear density was better for **4** than for **3**, as can be seen in the images below. Water molecules have banana shaped nuclear density because the deuteriums were visible clearly.



**Figure 38:** Neutron density map for **3** (left) showing the density for ligand (red) and key water molecules (blue). The water in question and other water molecules (black circle) were not visible in nuclear map, although it was visible in X-ray map. On right nuclear density map for key residues is shown. The  $2m|F_o| - D|F_c|$  map is contoured at  $1.0\sigma$ .

There was back exchange of the water molecules in the crystal for **3**. In **4**, the ligand and protein density were very good, but the water molecules showed poor nuclear density and lack of banana shape which is characteristic of water nuclear density map, as shown in the image below.



**Figure 39:** Nuclear density map for **4** showing clear density for ligand (red). The water in question and other water molecules (black circles) were not visible in the nuclear density map, although they were visible in the X-ray map. Right: nuclear density map for key residues. The  $2m|F_o| - D|F_c|$  map is contoured at  $1.0\sigma$ .

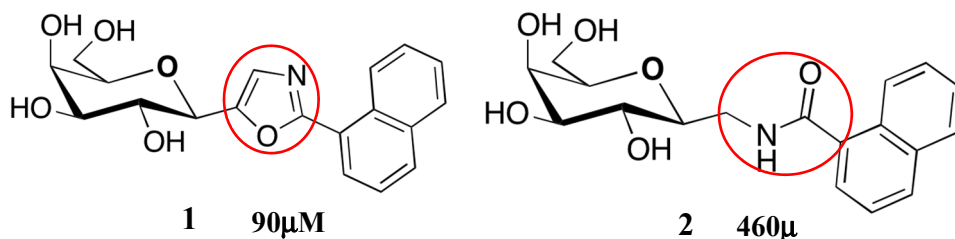
Unfortunately, the data were not good enough to analyse the water molecules and H-bonding question. For compound **3** there was simple back-exchange, which could have occurred during sample mounting, transport or data collection. For **4**, there is some other problem with the data, as the nuclear density did not show proper shape characteristics for water molecules. Hence, a re-collection of data for both the compound will be necessary to answer the question and analyse the binding.

**Refinement statistics**

compound	Resolution (Å) X-ray/neutron	X-ray $R_{\text{model}}/R_{\text{free}}$	Neutron $R_{\text{model}}/R_{\text{free}}$
3	1.30/1.85	0.154/0.168	0.190/0.232
4	1.15/1.80	0.122/0.140	0.142/0.186

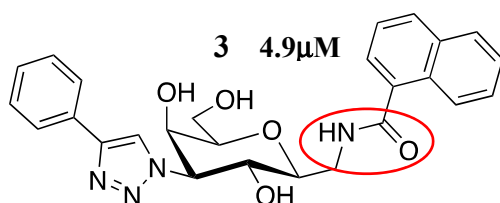
### Role of amide vs. oxazole group in ligand binding

The structures of two related monogalactoside- ligands are shown in Figure XX. Compound 1 has an oxazole group (red circle) attached to 1'-OH of galactose, while compound 2 has an amide group (red circle) attached.



Chemical structure of the two compounds. The chemical group of interest is highlighted with red circle.

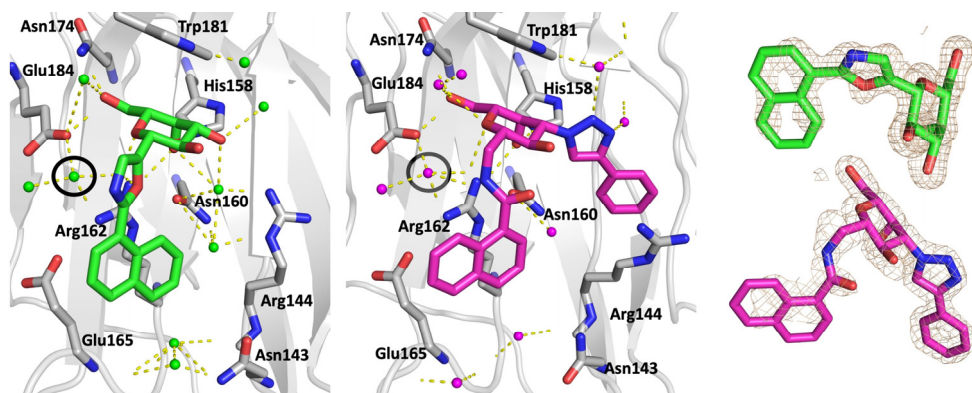
Binding affinities of these compounds were determined from FP. I solved 100K X-ray structures of these ligands in complex with the galectin-3 CRD.



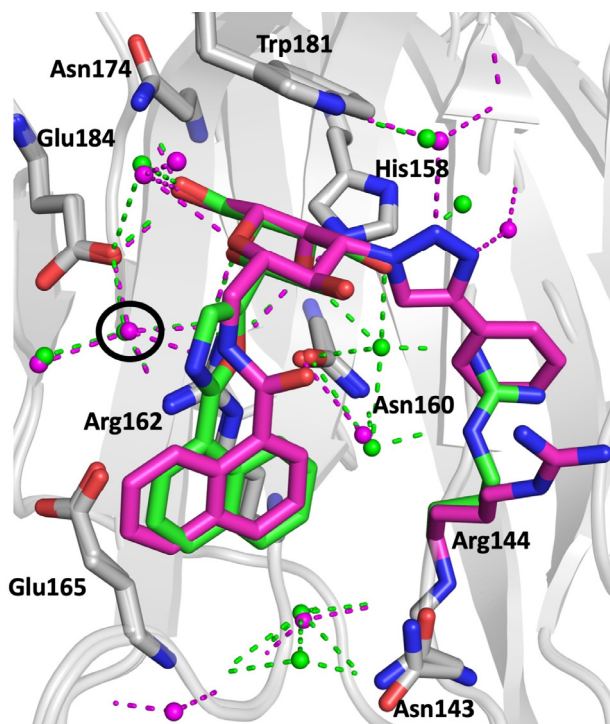
The compound 2 was modified to add a phenyltriazole group (3), which improved solubility and affinity as well.

### Results

The X-ray structures for **1** and **3** are compared below. Compound **1** and its key water molecules are coloured green and for compound **3** they are coloured magenta. All the water molecules that interact with the ligand are shown. The galactose moiety is well defined in both structures, while the naphthalene group has slightly poorer electron density. The phenyltriazole group in **3** is also well defined, and it interacts with Arg144. A water molecule of particular interest is circled in black. This water molecule sits between Glu184 and Arg162. The hypothesis is that the amide bond in **3** donates an H-bond to the water molecule in question. It is unclear if the oxazole group in **1** donates or accepts an H-bond, and if this can influence the relative affinities of the two compounds. Or is there a complete lack of H-bond for the oxazole group in **1**, which leads to different position compared to amide group in **3**?



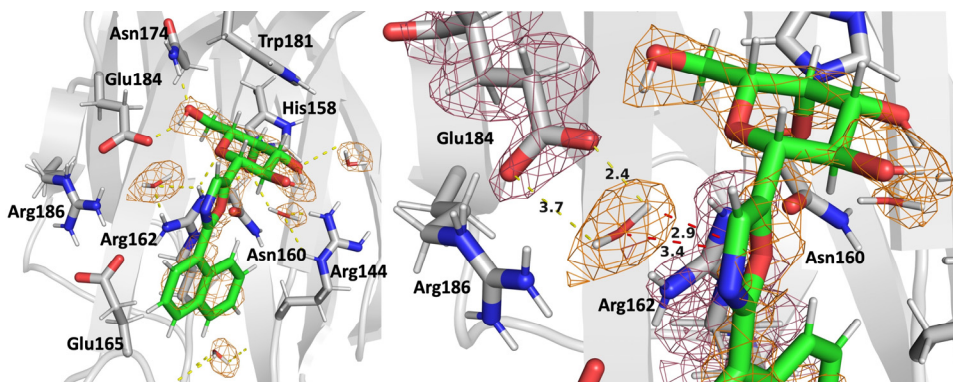
**Figure 40:** X-ray structures of **1** (green) and **3** (magenta) in complex with the galectin-3 CRD. The key water molecule in question, which was clearly visible in the structures, is highlighted by a black circle. Water molecules are coloured same as the ligands. The panel on the right depicts electron density of the ligands.



**Figure 41:** Comparison of the structures for **1** (green) and **3** (magenta). The water molecules are coloured similar to ligands to differentiate. As can be seen, the water molecules are well conserved in both the structures. Arg144 is moved by the phenyltriazole group in **3**. The key water molecule is circled.



To address the question, we needed neutron diffraction data for both the complexes. I prepared large crystals ( $> 1\text{ mm}^3$ ) of perdeuterated galectin-3 CRD and soaked them with 10 mM of compound **1**. The data were collected at the BioDiff instrument of the FRM-II facility in Munich, Germany. 47 minutes exposure time and  $0.4^\circ$  rotation per frame were used and 270 images were collected. Data were processed by beamline scientists using a modified version of HKL2000, and a scaled and merged data file was provided. The neutron data were of  $1.85\text{ \AA}$  resolution and the completeness was 90%.  $R_{\text{meas}}$  and  $R_{\text{pim}}$  were 0.21 and 0.12 respectively. A room temperature X-ray data set was collected on a different crystal soaked with **1**, as the crystal from neutron data collection dissolved during transport. The apo-galectin-3 CRD cryo structure was used as the model for joint refinement with water and hydrogens stripped. The X-ray data were of higher resolution ( $1.2\text{ \AA}$ ). Water molecules were added to the model based on nuclear density. The refined model was analyzed for the interaction of water molecule in question. As can be seen in the image below, the ligand and some of the water molecules showed good nuclear density. The water molecule in question had excellent density and showed the characteristic banana shape. In the right-hand panel, the distance of the oxygen in oxazole group to the deuterium in the water molecule is  $2.9\text{ \AA}$  and the oxygen atom is at  $3.4\text{ \AA}$ . The Glu165 side chain forms an H-bond with the water molecule, and the two deuteriums are  $2.4$  and  $3.7\text{ \AA}$  away. Although, the water is not aligned towards the Glu165 side chain, and oxazole group is also pointing away from the water.



**Figure 42:** Left: nuclear density for the ligand and water molecule (wheat colour). Right: close-up view of the water molecule, residues involved and oxazole moiety of ligand as well as showing important distances.

## Conclusion

Current analysis is incomplete as we need the neutron data for compound **3** to see complete picture. It is difficult to speculate anything conclusive from only one dataset. Although we expect to see the water molecule in a different binding mode compared to **1** and the amide mode is able to a H-bond as per our assumption.

## Refinement statistics

compound	Resolution (Å) X-ray/neutron	X-ray $R_{\text{model}}/R_{\text{free}}$	Neutron $R_{\text{model}}/R_{\text{free}}$
1 (cryo only)	1.01/N.A.	0.140/0.170	NA
1 (Joint_refinement)	1.19/1.85	0.139/0.165	0.260/0.310
3 (cryo only)	1.58/N.A.	0.150/0.193	N.A.

N.A.: not applicable





## 7. References

1. Blobaum, A. L. & Marnett, L. J. Structural and Functional Basis of Cyclooxygenase Inhibition. *J. Med. Chem.* **50**, 1425–1441 (2007).
2. Vane, J. R. & Botting, R. M. The mechanism of action of aspirin. *Thromb. Res.* **110**, 255–8 (2003).
3. Matsuyama, T., Yamashita, T., Imai, H. & Shichida, Y. Covalent bond between ligand and receptor required for efficient activation in rhodopsin. *J. Biol. Chem.* **285**, 8114–8121 (2010).
4. Zhang, C. & Lai, L. Towards structure-based protein drug design. *Biochem. Soc. Trans.* **39**, 1382–1386 (2012).
5. Meyer, E. A., Castellano, R. K. & Diederich, F. *Interactions with Arenes Interactions with Aromatic Rings in Chemical and Biological Recognition Angewandte. Angew. Chem. Int. Ed.* **42**, (2003).
6. Perozzo, R., Folkers, G. & Scapozza, L. Thermodynamics of protein-ligand interactions: History, presence, and future aspects. *J. Recept. Signal Transduct.* **24**, 1–52 (2004).
7. Klebe, G. *Applying thermodynamic profiling in lead finding and optimization. Nature Reviews Drug Discovery* **14**, (2015).
8. Olsson, T. S. G., Williams, M. A., Pitt, W. R. & Ladbury, J. E. The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **384**, 1002–1017 (2008).
9. Bissantz, C., Kuhn, B. & Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **53**, 6241–6241 (2010).
10. Olsson, T. S. G., Williams, M. A., Pitt, W. R. & Ladbury, J. E. The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design. *J. Mol. Biol.* **384**, 1002–1017 (2008).
11. Homans, S. W. Dynamics and thermodynamics of ligand-Protein interactions. *Top. Curr. Chem.* **272**, 51–82 (2007).
12. Ladbury, J. E., Klebe, G. & Freire, E. Adding calorimetric data to decision making in lead discovery: A hot tip. *Nat. Rev. Drug Discov.* **9**, 23–27 (2010).
13. Reynolds, C. H. & Holloway, M. K. Thermodynamics of ligand binding and efficiency. *ACS Med. Chem. Lett.* **2**, 433–437 (2011).

14. Chodera, J. D. & Mobley, D. L. Entropy-Enthalpy Compensation: Role and Ramifications in Biomolecular Ligand Recognition and Design. *Annu. Rev. Biophys.* **42**, 121–142 (2013).
15. Wiene-Schmidt, B. *et al.* Paradoxically, Most Flexible Ligand Binds Most Entropy-Favored: Intriguing Impact of Ligand Flexibility and Solvation on Drug-Kinase Binding. *J. Med. Chem.* **61**, 5922–5933 (2018).
16. Poornima, C. S. & Dean, P. M. *Hydration in drug design. 2. Influence of local site surface shape on water binding. Journal of Computer-Aided Molecular Design* **9**, (1995).
17. Dunitz, J. D. The entropic cost of bound water in crystals and biomolecules. *Science (80-. )*. **264**, 670 (1994).
18. Dubins, D. N., Filfil, R., Macgregor, R. B. & Chalikian, T. V. Role of Water in Protein–Ligand Interactions: Volumetric Characterization of the Binding of 2'-CMP and 3'-CMP to Ribonuclease A. *J. Phys. Chem. B* **104**, 390–401 (2002).
19. Duff, M. R. & Howell, E. E. Thermodynamics and solvent linkage of macromolecule-ligand interactions. *Methods* **76**, 51–60 (2015).
20. Schiebel, J. *et al.* Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. *Nat. Commun.* **9**, (2018).
21. Saraboji, K. *et al.* The carbohydrate-binding site in galectin-3 is preorganized to recognize a sugarlike framework of oxygens: Ultra-high-resolution structures and water dynamics. *Biochemistry* **51**, 296–306 (2012).
22. Krimmer, S. G. & Klebe, G. Thermodynamics of protein-ligand interactions as a reference for computational analysis: How to assess accuracy, reliability and relevance of experimental data. *J. Comput. Aided. Mol. Des.* **29**, 867–883 (2015).
23. Ferreira De Freitas, R. & Schapira, M. A systematic analysis of atomic protein-ligand interactions in the PDB. *Medchemcomm* **8**, 1970–1981 (2017).
24. Gohlke, H. & Klebe, G. *Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angewandte Chemie - International Edition* **41**, (2002).
25. BOHM, H.-J. & KLEBE, G. *What Can We Learn from Molecular Recognition in Protein-Ligand Complexes for the Design of New Drugs? Angewandte Chemie International Edition* **28**, (2010).
26. Steiner, T. The hydrogen bond in the solid state. *Angew. Chemie - Int. Ed.* **41**, 48–76 (2002).
27. Waldburger, C. D., Schildbach, J. F. & Sauer, R. T. Are buried salt bridges important for protein stability and conformational specificity? *Nat. Struct. Biol.* **2**, 122–128 (1995).
28. Gallivan, J. P. & Dougherty, D. A. A computational study of cation- $\pi$  interactions

- vs salt bridges in aqueous media: Implications for protein engineering. *J. Am. Chem. Soc.* **122**, 870–874 (2000).
29. Salonen, L. M. *et al.* Cation- $\pi$  interactions at the active site of factor Xa: Dramatic enhancement upon stepwise N-alkylation of ammonium ions. *Angew. Chemie - Int. Ed.* **48**, 811–814 (2009).
  30. Sörme, P. *et al.* Structural and thermodynamic studies on cation-II interactions in lectin-ligand complexes: High-affinity galectin-3 inhibitors through fine-tuning of an arginine-arene interaction. *J. Am. Chem. Soc.* **127**, 1737–1743 (2005).
  31. Auffinger, P., Hays, F. A., Westhof, E. & Ho, P. S. *Halogen bonds in biological molecules. Proceedings of the National Academy of Sciences* **101**, (2004).
  32. Parisini, E., Metrangolo, P., Pilati, T., Resnati, G. & Terraneo, G. Halogen bonding in halocarbon–protein complexes: a structural survey. *Chem. Soc. Rev.* **40**, 2267–2278 (2011).
  33. Zürcher, M. & Diederich, F. Structure-based drug design: Exploring the proper filling of apolar pockets at enzyme active sites. *J. Org. Chem.* **73**, 4345–4361 (2008).
  34. Müller, K., Faeh, C. & Diederich, F. *Fluorine in pharmaceuticals: Looking beyond intuition. Science* **317**, (2007).
  35. Pollock, J. *et al.* Rational Design of Orthogonal Multipolar Interactions with Fluorine in Protein-Ligand Complexes. *J. Med. Chem.* **58**, 7465–7474 (2015).
  36. Olsen, J. A. *et al.* A fluorine scan of thrombin inhibitors to map the fluorophilicity/fluorophobicity of an enzyme active site: Evidence for C-F...C=O interactions. *Angew. Chemie - Int. Ed.* **42**, 2507–2511 (2003).
  37. Olsen, J. A. *et al.* Fluorine interactions at the thrombin active site: Protein backbone fragments H-Ca-C=O comprise a favorable C-F environment and interactions of C-F with electrophiles. *ChemBioChem* **5**, 666–675 (2004).
  38. Giegé, R. A historical perspective on protein crystallization from 1840 to the present day. *FEBS J.* **280**, 6456–6497 (2013).
  39. Jaskolski, M., Dauter, Z. & Wlodawer, A. A brief history of macromolecular crystallography, illustrated by a family tree and its Nobel fruits. *FEBS J.* **281**, 3985–4009 (2014).
  40. Kendrew, J. C. *et al.* Structure of myoglobin: A three-dimensional fourier synthesis at 2 . resolution. *Nature* **185**, 422–427 (1960).
  41. Ursby, T. *et al.* The macromolecular crystallography beamline I911-3 at the MAX IV laboratory. *J. Synchrotron Radiat.* **20**, 648–653 (2013).
  42. Rhodes, G. *Crystallography made crystal clear: a guide for users of macromolecular models.* (Elsevier/Academic Press, 2006).
  43. Rupp, B. *Biomolecular crystallography: principles, practice, and application to structural biology.* (Garland Science, 2010).

44. Blundell, T. L. & Johnson, L. N. *Protein crystallography*. (Academic Press, 1976).
45. Hope, H. *Crystallography Of Biological Macromolecules At Ultra-Low Temperature. Annual Review of Biophysics and Biomolecular Structure* **19**, (1990).
46. Broennimann, C. *et al.* The PILATUS 1M detector. *J. Synchrotron Radiat.* **13**, 120–130 (2006).
47. Hasegawa, K. *et al.* Development of a shutterless continuous rotation method using an X-ray CMOS detector for protein crystallography. *J. Appl. Cryst* **42**, 1165–1175 (2009).
48. Karplus, P. A. & Diederichs, K. Assessing and maximizing data quality in macromolecular crystallography. *Curr. Opin. Struct. Biol.* **34**, 60–68 (2015).
49. Karplus, P. A. & Diederichs, K. Linking crystallographic model and data quality. *Science (80-. )*. **336**, 1030–1033 (2012).
50. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings IPII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) . *Adv. Drug Deliv. Rev.* **46**, 3–26 (2002).
51. van Montfort, R. L. M. & Workman, P. Structure-based drug design: aiming for a perfect fit. *Essays Biochem.* **61**, 431–437 (2017).
52. Blundell, T. L. Protein crystallography and drug discovery: recollections of knowledge exchange between academia and industry. *IUCrJ* **4**, 308–321 (2017).
53. Barondes, S. H. *et al.* Galectins: A family of animal  $\beta$ -galactoside-binding lectins. *Cell* **76**, 597–598 (1994).
54. Barondes, S. H., Cooper, D. N. W., Gitt, M. A. & Leffler, H. *Galectins. Structure and function of a large family of animal lectins. Journal of Biological Chemistry* **269**, (1994).
55. Johannes, L., Jacob, R. & Leffler, H. Galectins at a glance. *J. Cell Sci.* **131**, jcs208884 (2018).
56. Leffler, H., Carlsson, S., Hedlund, M., Qian, Y. & Poirier, F. *Introduction to galectins. Glycoconjugate Journal* **19**, (2004).
57. Lindstedt, R., Apodaca, G., Barondes, S. H., Mostov, K. E. & Leffler, H. *Apical secretion of a cytosolic protein by Madin-Darby Canine Kidney cells. Journal of biological chemistry* **268**, (1993).
58. Elola, M. T., Wolfenstein-Todel, C., Troncoso, M. F., Vasta, G. R. & Rabinovich, G. A. Galectins: Matricellular glycan-binding proteins linking cell adhesion, migration, and survival. *Cell. Mol. Life Sci.* **64**, 1679–1700 (2007).
59. Cummings, R. D., Liu, F.-T. & Vasta, G. R. *Galectins. Essentials of Glycobiology* **5**, (2015).

60. Liao, D.-I., Kapadia, G., Ahmedt, H., Vastat, G. R. & Herzberg, O. *Structure of S-lectin, a developmentally regulated vertebrate , $\beta$ -galactoside-binding protein. Proc. Nati. Acad. Sci. USA* **91**, (1994).
61. Seetharaman, J. *et al.* X-ray crystal structure of the human galectin-3 carbohydrate recognition domain at 2.1-angstrom resolution. *J. Biol. Chem.* **273**, 13047–13052 (1998).
62. Liu, F. T. & Rabinovich, G. A. Galectins as modulators of tumour progression. *Nature Reviews Cancer* **5**, 29–41 (2005).
63. Kim, B. W., Beom Hong, S., Hoe Kim, J., Hoon Kwon, D. & Song, H. K. Structural basis for recognition of autophagic receptor NDP52 by the sugar receptor galectin-8. *Nat. Commun.* **4**, (2013).
64. Li, S. *et al.* Sterical hindrance promotes selectivity of the autophagy cargo receptor NDP52 for the danger receptor galectin-8 in antibacterial autophagy. *Sci. Signal.* **6**, (2013).
65. Sacchettini, J. C., Baum, L. G. & Brewer, C. F. Multivalent protein - Carbohydrate interactions. A new paradigm for supermolecular assembly and signal transduction. *Biochemistry* **40**, 3009–3015 (2001).
66. Fred Brewer, C. *Binding and cross-linking properties of galectins. Biochimica et Biophysica Acta - General Subjects* **1572**, (2002).
67. Hirabayashi, J. *et al.* *Oligosaccharide specificity of galectins: a search by frontal affinity chromatography. Biochimica et Biophysica Acta* **1572**, (2002).
68. Wilson, T. J., Firth, M. N., Powell, J. T. & Harrison, F. L. The sequence of the mouse 14 kDa  $\beta$ -galactoside-binding lectin and evidence for its synthesis on free cytoplasmic ribosomes. *Biochem. J.* **261**, 847–52 (1989).
69. Cooper D. N., B. S. H. *Evidence for export of a muscle lectin from cytosol to extracellular matrix and for a novel secretory mechanism. The Journal of Cell Biology* **110**, (1990).
70. Nickel, W. The mystery of nonclassical protein secretion. *Eur. J. Biochem.* **270**, 2109–2119 (2003).
71. Hughes, R. C. *Galectins as modulators of cell adhesion. Biochimie* **83**, (2001).
72. Hikita, C. *et al.* *Induction of terminal differentiation in epithelial cells requires polymerization of hensin by galectin 3. Journal of Cell Biology* **151**, (2000).
73. Dagher, S. F., Wang, J. L. & Patterson, R. J. *Identification of galectin-3 as a factor in pre-mRNA splicing. Proceedings of the National Academy of Sciences of the United States of America* **92**, (1995).
74. Liu, F. T., Patterson, R. J. & Wang, J. L. *Intracellular functions of galectins. Biochimica et Biophysica Acta - General Subjects* **1572**, (2002).
75. Ahmad, N. *et al.* Galectin-3 Precipitates as a Pentamer with Synthetic Multivalent

- Carbohydrates and Forms Heterogeneous Cross-linked Complexes\*. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M312834200
76. Lepur, A., Salomonsson, E., Nilsson, U. J. & Leffler, H. Ligand induced galectin-3 protein self-association. *J. Biol. Chem.* **287**, 21751–21756 (2012).
  77. Ho, M. K. & Springer, T. A. Mac-2, a novel 32,000 Mr mouse macrophage subpopulation-specific antigen defined by monoclonal antibodies. *J. Immunol.* **128**, 1221–1228 (1982).
  78. Chiariotti, L., Salvatore, P., Frunzio, R. & Bruni, C. B. *Galectin genes: Regulation of expression*. *Glycoconjugate Journal* **19**, (Kluwer Academic Publishers, 2002).
  79. Dumić, J., Dabelić, S. & Flögel, M. Galectin-3: An open-ended story. *Biochim. Biophys. Acta - Gen. Subj.* **1760**, 616–635 (2006).
  80. Takenaka, Y., Fukumori, T. & Raz, A. *Galectin-3 and metastasis*. *Glycoconjugate Journal* **19**, (Kluwer Academic Publishers, 2002).
  81. Newlaczyl, A. U. & Yu, L. G. Galectin-3 - A jack-of-all-trades in cancer. *Cancer Lett.* **313**, 123–128 (2011).
  82. Sciacchitano, S. *et al.* Galectin-3: One molecule for an alphabet of diseases, from A to Z. *Int. J. Mol. Sci.* **19**, (2018).
  83. Gong, H. C. *et al.* The NH<sub>2</sub> terminus of galectin-3 governs cellular compartmentalization and functions in cancer cells. *Cancer Research* **59**, (1999).
  84. Birdsall, B. *et al.* NMR solution studies of hamster galectin-3 and electron microscopic visualization of surface-adsorbed complexes: Evidence for interactions between the N- and C-terminal domains. *Biochemistry* **40**, 4859–4866 (2001).
  85. Barboni, E. A. M., Bawumia, S. & Hughes, R. C. Kinetic measurements of binding of galectin 3 to a laminin substratum. *Glycoconjugate Journal* **16**, (1999).
  86. Raz, A., Pazerini, G. & Carmi, P. Identification of the Metastasis-associated, Galactoside-binding Lectin as a Chimeric Gene Product with Homology to an IgE-binding Protein. *Cancer Research* **49**, (1989).
  87. Raz, A. S. N.-M. P. I. H. K. H. *Galectin-3: A Novel Antiapoptotic Molecule with A Functional BHL (NWGR) Domain of Bcl-2 Family*. *Karmanos Cancer institute* **1**, (1997).
  88. Yang, R.-Y., Hill, P. N., Hsu, D. K. & Liu, F.-T. Role of the Carboxyl-Terminal Lectin Domain in Self-Association of Galectin-3. *Biochemistry* **37**, (1998).
  89. Yoshii, T. *et al.* Galectin-3 phosphorylation is required for its anti-apoptotic function and cell cycle arrest. *J. Biol. Chem.* **277**, 6852–6857 (2002).
  90. Berbís, M. Á. *et al.* Peptides derived from human galectin-3 N-terminal tail interact with its carbohydrate recognition domain in a phosphorylation-dependent manner. *Biochemical and Biophysical Research Communications* **443**, (Elsevier, 2014).

91. Ippel, H. *et al.* Intra- and intermolecular interactions of human galectin-3: Assessment by full-assignment-based NMR. *Glycobiology* **26**, 888–903 (2016).
92. Lin, Y. H. *et al.* The intrinsically disordered N-terminal domain of galectin-3 dynamically mediates multisite self-association of the protein through fuzzy interactions. *J. Biol. Chem.* **292**, 17845–17856 (2017).
93. Flores-Ibarra, A., Vértessy, S., Medrano, F. J., Gabius, H. J. & Romero, A. Crystallization of a human galectin-3 variant with two ordered segments in the shortened N-terminal tail. *Sci. Rep.* **8**, (2018).
94. Menon, R. P. & Hughes, R. C. *Determinants in the N-terminal domains of galectin-3 for secretion by a novel pathway circumventing the endoplasmic reticulum-Golgi complex. European Journal of Biochemistry* **264**, (1999).
95. Ochieng, J., Green, B., Evans, S., James, O. & Warfield, P. *Modulation of the biological functions of galectin-3 by matrix metalloproteinases. Biochimica et Biophysica Acta - General Subjects* **1379**, (1998).
96. Leffler, H. & Barondes, S. H. *Specificity of binding of three soluble rat lung lectins to substituted and unsubstituted mammalian ??-galactosides. Journal of Biological Chemistry* **261**, (1986).
97. Bachhawat-Sikder, K., Thomas, C. J. & Surolia, A. *Thermodynamic analysis of the binding of galactose and poly-N-acetyllactosamine derivatives to human galectin-3. FEBS Letters* **500**, (2001).
98. Wang, J. L., Gray, R. M., Haudek, K. C. & Patterson, R. J. Nucleocytoplasmic lectins. *Biochim. Biophys. Acta - Gen. Subj.* **1673**, 75–93 (2004).
99. Paron, I. *et al.* Nuclear localization of Galectin-3 in transformed thyroid cells: A role in transcriptional regulation. *Biochem. Biophys. Res. Commun.* **302**, 545–553 (2003).
100. Nakahara, S. *et al.* Characterization of the nuclear import pathways of galectin-3. *Cancer Res.* **66**, 9995–10006 (2006).
101. Haudek, K. C. *et al.* Dynamics of galectin-3 in the nucleus and cytoplasm. *Biochim. Biophys. Acta - Gen. Subj.* **1800**, 181–189 (2010).
102. Yu, F., Finley, R. L., Raz, A. & Kim, H. R. C. Galectin-3 translocates to the perinuclear membranes and inhibits cytochrome c release from the mitochondria. A role for synexin in galectin-3 translocation. *J. Biol. Chem.* **277**, 15819–15827 (2002).
103. Elad-Sfadia, G., Haklai, R., Balan, E. & Kloog, Y. Galectin-3 augments K-ras activation and triggers a ras signal that attenuates ERK but not phosphoinositide 3-kinase activity. *J. Biol. Chem.* **279**, 34922–34930 (2004).
104. Oka, N. *et al.* Galectin-3 inhibits tumor necrosis factor-related apoptosis-inducing ligand-induced apoptosis by activating Akt in human bladder carcinoma cells. *Cancer Res.* **65**, 7546–7553 (2005).



105. Ruvolo, P. P. Galectin 3 as a guardian of the tumor microenvironment. *Biochim. Biophys. Acta - Mol. Cell Res.* **1863**, 427–437 (2016).
106. Patterson, R. J., Wang, W. & Wang, J. L. *Understanding the biochemical activities of galectin-1 and galectin-3 in the nucleus.* *Glycoconjugate Journal* **19**, (Kluwer Academic Publishers, 2002).
107. Shimura, T. *et al.* *Advances in Brief Galectin-3 , a Novel Binding Partner of  $\beta$ -Catenin.* *Cancer Research* **64**, (2004).
108. Shi, Y. *et al.* Inhibition of Wnt-2 and galectin-3 synergistically destabilizes  $\beta$ -catenin and induces apoptosis in human colorectal cancer cells. *Int. J. Cancer* **121**, 1175–1181 (2007).
109. Liu, F. T. & Rabinovich, G. A. Galectins: regulators of acute and chronic inflammation. *Year Immunol.* **2** **1183**, 158–182 (2010).
110. Ochieng, J., Furtak, V. & Lukyanov, P. *Extracellular functions of galectin-3.* *Glycoconjugate Journal* **19**, (Kluwer Academic Publishers, 2002).
111. Furtak, V., Hatcher, F. & Ochieng, J. Galectin-3 mediates the endocytosis of  $\alpha$ -1 integrins by breast carcinoma cells. *Biochem. Biophys. Res. Commun.* **289**, 845–850 (2001).
112. Diehl, C. *et al.* Protein flexibility and conformational entropy in ligand design targeting the carbohydrate recognition domain of galectin-3. *J. Am. Chem. Soc.* **132**, 14577–14589 (2010).
113. Collins, P. M., Öberg, C. T., Leffler, H., Nilsson, U. J. & Blanchard, H. Taloside Inhibitors of Galectin-1 and Galectin-3. *Chem. Biol. Drug Des.* **79**, 339–346 (2012).
114. Bum-Erdene, K. *et al.* Investigation into the feasibility of thioditaloside as a novel scaffold for galectin-3-specific inhibitors. *ChemBioChem* **14**, 1331–1342 (2013).
115. Cumpstey, I., Carlsson, S., Leffler, H. & Nilsson, U. J. Synthesis of a phenyl thio- $\beta$ -D-galactopyranoside library from 1,5-difluoro-2,4-dinitrobenzene: Discovery of efficient and selective monosaccharide inhibitors of galectin-7. *Org. Biomol. Chem.* **3**, 1922–1932 (2005).
116. Sörme, P., Kahl-Knutsson, B., Huflejt, M., Nilsson, U. J. & Leffler, H. Fluorescence polarization as an analytical tool to evaluate galectin-ligand interactions. *Anal. Biochem.* **334**, 36–47 (2004).
117. Zhuang, T., Leffler, H. & Prestegard, J. H. Enhancement of bound-state residual dipolar couplings: Conformational analysis of lactose bound to Galectin-3. *Protein Sci.* **15**, 1780–1790 (2006).
118. Leffler, H., Masiarz, F. R. & Barondes, S. H. Soluble Lactose-Binding Vertebrate Lectins: A Growing Family. *Biochemistry* **28**, 9222–9229 (1989).
119. Kabsch, W. *XDS.* in *Acta Crystallographica Section D Biological Crystallography* (eds. Rossmann, M. G. & Arnold, E.) **66**, 125–132 (Kluwer Academic Publishers,

- 2010).
120. Assmann, G., Brehm, W. & Diederichs, K. Identification of rogue datasets in serial crystallography. *J. Appl. Crystallogr.* **49**, 1021–1028 (2016).
  121. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. Sect. D Biol. Crystallogr.* **69**, 1204–1214 (2013).
  122. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2012). doi:10.1107/S0907444912001308
  123. Adams, P. D. *et al.* The Phenix software for automated determination of macromolecular structures. *Methods* (2011). doi:10.1016/j.ymeth.2011.07.005
  124. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D-Biological Crystallogr.* **60**, 2126–2132 (2004).
  125. Moriarty, N. W., Grosse-Kunstleve, R. W. & Adams, P. D. Electronic ligand builder and optimization workbench (eLBOW): A tool for ligand coordinate and restraint generation. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **65**, 1074–1080 (2009).
  126. Long, F. *et al.* AceDRG : a stereochemical description generator for ligands . *Acta Crystallogr. Sect. D Struct. Biol.* **73**, 112–122 (2017).
  127. Kuriyan, J. & Weis, W. I. Rigid protein motion as a model for crystallographic temperature factors. *Proc. Natl. Acad. Sci. U. S. A.* **88**, 2773–7 (1991).
  128. Schomaker, V. & Trueblood, K. N. On the rigid-body motion of molecules in crystals. *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **24**, 63–76 (2002).
  129. Trueblood, K. N. *et al.* Atomic displacement parameter nomenclature report of a subcommittee on atomic displacement parameter nomenclature. *Acta Crystallogr. Sect. A Found. Crystallogr.* **52**, 770–781 (1996).
  130. Joosten, R. P., Long, F., Murshudov, G. N. & Perrakis, A. The PDB\_REDO server for macromolecular structure model optimization . *IUCrJ* **1**, 213–220 (2014).
  131. Oksanen, E., Chen, J. C. H. & Fisher, S. Z. Neutron crystallography for the study of hydrogen bonds in macromolecules. *Molecules* **22**, (2017).
  132. Hirano, Y., Takeda, K. & Miki, K. Charge-density analysis of an iron-sulfur protein at an ultra-high resolution of 0.48 Å. *Nature* **534**, 281–284 (2016).
  133. Blakeley, M. P., Langan, P., Niimura, N. & Podjarny, A. Neutron crystallography: opportunities, challenges, and limitations. *Curr. Opin. Struct. Biol.* **18**, 593–600 (2008).
  134. Niimura, N. & Bau, R. Neutron protein crystallography: beyond the folding structure of biological macromolecules. *Acta Crystallogr. Sect. A Found. Crystallogr.* **64**, 12–22 (2008).

135. Kwon, H., Langan, P. S., Coates, L., Raven, E. L. & Moody, P. C. E. The rise of neutron cryo-crystallography. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 792–799 (2018).
136. O'Dell, W. B., Bodenheimer, A. M. & Meilleur, F. Neutron protein crystallography: A complementary tool for locating hydrogens in proteins. *Arch. Biochem. Biophys.* **602**, 48–60 (2016).
137. Blakeley, M. P., Hasnain, S. S. & Antonyuk, S. V. Sub-atomic resolution X-ray crystallography and neutron crystallography: promise, challenges and potential. *IUCrJ* **2**, 464–474 (2015).
138. Blakeley, M. P. Neutron macromolecular crystallography. *Crystallogr. Rev.* **15**, 157–218 (2009).
139. Meilleur, F., Myles, D. A. A. & Blakeley, M. P. Neutron Laue macromolecular crystallography. *Eur. Biophys. J.* **35**, 611–620 (2006).
140. O'Dell, W. B., Bodenheimer, A. M. & Meilleur, F. Neutron protein crystallography: A complementary tool for locating hydrogens in proteins. *Arch. Biochem. Biophys.* **602**, 48–60 (2016).
141. Blakeley, M. P. *et al.* Neutron macromolecular crystallography with LADI-III. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 1198–1205 (2010).
142. Ostermann, A. & Schrader, T. BIODIFF: Diffractometer for large unit cells. *J. large-scale Res. Facil. JLSRF* **1**, A2 (2015).
143. Coates, L. *et al.* The Macromolecular Neutron Diffractometer MaNDi at the Spallation Neutron Source. *J. Appl. Crystallogr.* **48**, 1302–1306 (2015).
144. Niimura, N. *et al.* Neutron Laue diffractometry with an imaging plate provides an effective data collection for neutron protein crystallography. *Phys. B Condens. Matter* **241–243**, 1162–1165 (1997).
145. Campbell, J. W., Hao, Q., Harding, M. M., Nguti, N. D. & Wilkinson, C. LAUEGEN version 6.0 and INTLDM. *J. Appl. Crystallogr.* **31**, 496–502 (1998).
146. Liebschner, D., Afonine, P. V., Moriarty, N. W., Langan, P. & Adams, P. D. Evaluation of models determined by neutron diffraction and proposed improvements to their validation and deposition. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 800–813 (2018).
147. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2010). doi:10.1107/S0907444909052925
148. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with phenix.refine. **68**, 352–367 (2012).
149. Afonine, P. V. *et al.* Joint X-ray and neutron refinement with phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 1153–1163 (2010).

150. Zetterberg, F. R. *et al.* Monosaccharide Derivatives with Low-Nanomolar Lectin Affinity and High Selectivity Based on Combined Fluorine–Amide, Phenyl–Arginine, Sulfur– $\pi$ , and Halogen Bond Interactions. *ChemMedChem* **13**, 133–137 (2018).
151. Noresson, A. L. *et al.* Designing interactions by control of protein-ligand complex conformation: Tuning arginine-arene interaction geometry for enhanced electrostatic protein-ligand interactions. *Chem. Sci.* **9**, 1014–1021 (2018).





