

A HIGH-THROUGHPUT COMPUTATIONAL MODEL FOR PREDICTION AND MITIGATION OF BIO-THERAPEUTIC PROTEIN AGGREGATION

by

RIT PRATIK MISHRA



DEPARTMENT OF CHEMICAL
ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
DELHI
DECEMBER 2023

A HIGH-THROUGHPUT COMPUTATIONAL MODEL FOR PREDICTION AND MITIGATION OF BIO-THERAPEUTIC PROTEIN AGGREGATION

by

RIT PRATIK MISHRA
DEPARTMENT OF CHEMICAL
ENGINEERING

*Submitted in fulfillment of requirements of degree of
Doctor of Philosophy*



INDIAN INSTITUTE OF TECHNOLOGY
DELHI

December 2023

Dedicated to,

*My beloved family, without whose endless love
and support, I could not achieve this.*

CERTIFICATE

This is to certify that the thesis entitled "A High-Throughput Computational Model for Prediction and Mitigation of Bio-Therapeutic Protein Aggregation", submitted by Mr. Rit Pratik Mishra to the Indian Institute of Technology Delhi for the award of the degree of Doctor of Philosophy, is a record of the original bonafide research work carried out by him. He has worked under my supervision and has fulfilled the requirements, which to my knowledge, has reached the requisite standard for the submission of this thesis. The results in this thesis have not been submitted in part or whole to any University or institute to award any degree or diploma.

Date: **27.12.2023**

Place: **Delhi**



Prof. Gaurav Goel

Professor

Department of Chemical Engineering

Indian Institute of Technology Delhi

New Delhi, India

Acknowledgments

I would like to express my sincere gratitude and appreciation to the following individuals and institutions for their invaluable support and contributions throughout the course of my research. First and foremost, I am deeply grateful to my supervisor, Professor Gaurav Goel, for his unwavering guidance, patience, and expertise. His insightful feedback, continuous encouragement, and dedication to my academic growth have been instrumental in shaping this thesis. I would like to extend my heartfelt thanks to Professor Anurag Singh Rathore, who served as the Chairman of my review committee. His valuable insights, critical evaluation, and constructive suggestions have greatly enriched the quality of this research work. I am also indebted to Professor Sudip Kumar Pattanayak, the department expert in my committee, for his valuable inputs, discussions, and guidance throughout this journey. His expertise and thorough review of my work have significantly contributed to its improvement. I extend my gratitude to Professor Tapan Kumar Chaudhuri, the external department expert, for his insightful comments and valuable suggestions during the evaluation process. His expertise and feedback have been immensely valuable in shaping the outcomes of this research.

I am deeply indebted to my family: my father, Rajendra Kumar Mishra, my mother, Gitanjali Mishra, and my brother, Rajpratik Mishra. Their unwavering love, support, and understanding have been the pillars of my strength, motivation, and perseverance throughout this endeavor. I am grateful to my dear friends, and my fiancé Rinnie, for their constant support, motivation, and unwavering presence throughout this journey. Their friendship, love, and encouragement have been invaluable in navigating the challenges encountered during this research. I would like to express my gratitude to my seniors, Richa Singh Rathore and Avinash Mishra, for their valuable advice, encouragement, and willingness to share their knowledge and experiences. Their insights have been invaluable in navigating the challenges encountered during this research.

I would like to acknowledge the financial support I received from the MHRD scholarship provided by the Indian Institute of Technology (IIT) Delhi. This scholarship played a crucial role in supporting my research and enabling me to pursue my academic goals. Additionally, I would like to express my gratitude to the HPC facility at IIT Delhi for granting me access to their high-performance computing resources. The computational resources provided by the facility were instrumental in performing complex simulations and data analysis, enhancing the quality and depth of my research. I would also like to extend my appreciation to our lab assistant, Krishna Kumar, for his invaluable assistance and support throughout my research. His technical expertise, willingness to help, and efficient management of laboratory resources have greatly facilitated my experiments and data collection.

To all those mentioned above and to anyone else who has contributed in ways big or small, I offer my sincerest thanks. This thesis would not have been possible without your support, encouragement, and belief in my abilities.

A special mention goes to my pet cat, Billoo, who provided unconditional love and companionship during the challenging times of the COVID-19 lockdown in Delhi. His presence brought comfort and solace, making the difficult days more bearable. Lastly, I would like to express my gratitude to Lord Hanuman, whose divine blessings and guidance have provided me with strength and determination, especially during times of self-doubt and difficulty. His light has shown me the way and kept my spirits high.

Date: December 2023

Place: Bangalore



Rit Pratik Mishra
Roll No.: 2016CHZ8292

Abstract

Protein aggregation has been identified as one of the main problems in the manufacture and storage of biotherapeutics. Mechanistic understanding and quantitative predictions of aggregation kinetics would enable design of effective mitigation strategies. A comprehensive understanding of the aggregation process at the molecular level is essential to gain better structural insights and a quantitative understanding of protein-protein interactions (PPIs) and to predict kinetic parameters involved in each stage (folding/unfolding, nucleation, and fibrillation). However, studying the entire process at a single resolution is not feasible due to the different time and length scales at which these stages occur. In the present thesis, we have developed a multiscale computational model to accurately predict protein aggregation kinetics parameters, provide insights into molecular determinants of PPIs involved in aggravating aggregation, and quantify the effects of various stress factors (monomer concentration, excipients, temperature) on the aggregation kinetics. Two representative biotherapeutic proteins were studied: insulin, a small globular protein, and rituximab, an IgG1 antibody, as representative biotherapeutic proteins. These proteins are of commercial significance and necessitate stable formulation designs. However, due to the substantial disparity in size (with rituximab being approximately 25 times larger than insulin), computational complexities arise, particularly in terms of computational resource requirements and achieving convergence in sampling. The designed predictive model was used successfully to make accurate predictions for insulin aggregation kinetics and discover FDA approved inactives to inhibit its aggregation. Finally, we developed a protocol to sample aggregation prone regions (APRs) of rituximab, which can be extended to other classes of monoclonal antibodies.

Even at its lowest complexity, understanding protein aggregation requires comprehensive knowledge of single molecule conformational flexibility and bio-molecular assembly over a broad range of timescales. For insulin and mAbs, this spans timescales on the order of femtoseconds (bond vibrations) to seconds and hours (aggregation nucleation). We combined kinetic, thermodynamic, and structural information from single molecule (protein folding) and two molecules (association) explicit-solvent simulations for the determination of kinetic parameters in protein aggregation nucleation with insulin as a model protein. We observed this "early-events" enough to accurately predict the kinetic parameters up to the nucleation event. A structural bioinformatics approach was developed to account for heterogeneity of aggregation-prone species with the transition complex theory applicable

in modelling association kinetics involving non-native species. Specifically, the kinetic pathway for formation of aggregation-prone oligomeric species was found to contain a structurally specific dominant binding mode, making the kinetic process similar to native protein association. The kinetic parameters thus obtained were used in a population balance model and accurate predictions for aggregation nucleation time varying over two orders of magnitude with changes in either insulin concentration or an aggregation-inhibitor ligand concentration were obtained, while an empirical parameter set was not found to be transferable for prediction of ligand effects. Further, this physically determined kinetic parameter set provided several mechanistic insights, such as the identification of the rate-limiting step in aggregation nucleation and a quantitative explanation for the switch from Arrhenius to non-Arrhenius aggregation kinetics around the melting temperature of insulin.

Establishing structural and interaction specificity in insulin self-association along the non-native aggregation pathway enabled us to use structure-based, computer aided methods for identification of ligands with high-specificity for inhibiting protein-protein interactions (PPIs) implicated in early-stage aggregation nucleation. We proceeded to identify lead molecules from the FDA inactives database that can inhibit insulin aggregation using a multi-level virtual screening protocol. Some key challenges specific to the present problem, when compared with native protein association, include structural heterogeneity of the protein species involved, multiple association pathways, and relatively higher probability of conformational rearrangement of the association complex. A first-level screening of the inactives database was based on dominant pharmacophore features of three short peptides and an organic fluorogen, both shown to significantly inhibit insulin aggregation nucleation by binding to aggregation-prone conformations of insulin. We then performed ensemble docking of various low-energy ligand conformations on aggregation-prone conformers of insulin followed by molecular dynamics simulations and binding affinity calculations on a subset of docked complexes to identify a subset of lead molecules with a high probability for inhibiting insulin aggregation nucleation. We then applied our designed aggregation nucleation kinetics model to account for protein-ligand interaction details and conformational rearrangement effects during protein oligomerization to determine ligand concentration for effective aggregation inhibition. The effect of these select ligands (lead compounds) aggregation inhibition was extensively investigated by incubating insulin at various concentrations under aggregation-prone aqueous buffer conditions (low pH, high temperature). Herein, aggregation kinetics were quantified using size exclusion chromatography and the secondary structure was determined using circular dichroism spectroscopy to quantify effect of ligand on native structure content after co-incubation with insulin under aggregation-prone conditions. Specifically, two lead compounds (Riboflavin and Aspartame) showed significant retardation of insulin aggregation at formulation concentration when used at a 2:1 ligand: protein molar concentration.

The second biotherapeutic of interest, rituximab, is a IgG1 monoclonal antibody used in the treat chronic lymphocytic leukemia. Linear response driven

molecular dynamics simulations, based on accelerating conformational transition along the slow normal modes, were used to efficiently and extensively explore the configuration space, and thus, identify partially-folded aggregation-prone conformations of rituximab at acidic conditions. A structural bioinformatic protocol based on combining information on residue hydrophobicity, solvent accessible surface area, β -propensity, and protein tertiary structure (3D structure) was used to identify potential aggregation hotspots in these partially-folded rituximab conformers. A set of putative non-native dimeric complexes, totaling 57 complexes, was generated through a multi-step approach involving rigid body docking with APRs as the target and subsequent molecular dynamics simulations for conformational rearrangement. MMPBSA calculations were performed on these complexes to evaluate the contribution of APRs to the non-native binding free energy. Based on the ranking obtained from these calculations, we identified nine critical APRs (six in Fc, two in Fab, and one in Hinge) that play a key role in driving the aggregation of rituximab. The APRs identified in the Fc region were explicitly localized within the C_{H2} domain, which aligns with the findings from Differential scanning calorimetry (DSC) investigations conducted on IgG1 antibodies under different pH conditions. Additionally, the presence of the identified APR in the hinge region has already been substantiated through mutation studies. Furthermore, the APRs in the Fab region were found to reside within the complementarity-determining region (CDR) loop, with their sequences overlapping with residues involved in antigen binding. These findings provide valuable insights into the underlying mechanisms of rituximab aggregation and its potential implications in bio-pharmaceuticals development.

सार

प्रोटीन एकत्रिति को निर्माण और भंडारण में मुख्य समस्याओं में से एक के रूप में पहचाना गया है। एकैकिक अवलोकन और मात्रात्मक पूर्वानुमान की मृदुता के साथ एकैकिक अवलोकन ने प्रोटीन-प्रोटीन संवादों में शामिल एकद्वितीयता के बेहतर संरचना अंदाज और प्रत्येक चरण (फोल्डिंग/अनफोल्डिंग, न्यूक्लियेशन, और फाइब्रिलेशन) में शामिल होने वाले किनेटिक पैरामीटरों की यथात्मक समझ प्राप्त करने की संवादिक समझ होना आवश्यक है। हालांकि, इन चरणों का एक ही संकलन में अध्ययन करना योग्य नहीं है क्योंकि इन चरणों को प्राप्त करने के लिए अलग-अलग समय और लंबाई की स्थायिता होती है। वर्तमान थीसिस में, हमने प्रोटीन एकत्रिति की गतिकी पैरामीटरों का सटीक अनुमान लगाने, एकत्रिति को बढ़ाने वाले प्रोटीन-प्रोटीन संवादों के मौलिक निर्धारकों के बारे में अंदाजा देने, और विभिन्न तनाव कारकों (मोनोमर संघटन, अल्पका, तापमान) के प्रभाव को न्यूनतम करने के प्रभाव को मात्रात्मक रूप से मापन करने के लिए एक बहुमानक संगणन मॉडल विकसित किया है। दो प्रतिनिधित्वकर्ता बायोथेरेप्यूटिक प्रोटीनों का अध्ययन किया गया: इंसुलिन, एक छोटे गोल मौलिक प्रोटीन, और रिटक्सिमैब, एक IgG1 एंटीबॉडी, जो व्यापारिक महत्व के हैं और स्थिर रूप से निर्माण डिजाइन की आवश्यकता है। हालांकि, आकार में विशाल अंतर (रिटक्सिमैब लगभग इंसुलिन के 25 गुना बड़ा होने के बावजूद), विशेष रूप से संगणन संसाधन आवश्यकताओं और सैम्पलिंग में संघटन प्राप्त करने के प्रक्रियाकलित मुद्दे उत्पन्न होते हैं। डिजाइन किया गया पूर्वानुमानित मॉडल सफलतापूर्वक इंसुलिन एकत्रिति की गतिकी पैरामीटरों के लिए सटीक पूर्वानुमान बनाने के लिए उपयोग किया गया और उसकी एकत्रिति को रोकने के लिए FDA द्वारा मंजूर किए गए असक्रिय तत्वों की खोज की। आखिरकार, हमने रिटक्सिमैब के एकत्रिति प्रवृत्ति क्षेत्रों (एएपीआर) को सैम्पल करने के लिए प्रोटोकॉल विकसित किया, जो अन्य प्रकार की मोनोक्लोनल एंटीबॉडियों के लिए बढ़ाया जा सकता है।

प्रोटीन एकत्रिति को समझने के लिए उसके सबसे न्यून संघटन में भी, एक मोलेक्यूल के आकार बदलने की लचीलापन और बाइओ-मोलेक्यूलर समूह को बड़े संविदान घड़ी के स्तर पर जानकारी की आवश्यकता होती है। इंसुलिन और mAbs के लिए, इसमें फेम्टोसेकंड्स (बॉन्ड वाइब्रेशन) से सेकंड और घंटों (एग्रीगेशन न्यूक्लियेशन) के आदेश के कार्यकाल परिसर होते हैं। हमने इंसुलिन को मॉडल प्रोटीन के रूप में प्रोटीन एग्रीगेशन न्यूक्लियेशन में किनेटिक पैरामीटर्स की निर्धारण के लिए सिंगल मोलेक्यूल (प्रोटीन फोल्डिंग) और दो मोलेक्यूल (संघटन) स्पष्ट-सोल्वेंट सिमुलेशन से किनेटिक, थर्मोडायनमिक, और संरचनात्मक जानकारी को मिलाया। हमने इस "प्रारंभिक घटनाओं" को काफी समय तक देखा ताकि हम न्यूक्लियेशन घटना तक किनेटिक पैरामीटर्स की सटीक पूर्वानुमान कर सकें। हमने एक संरचनात्मक जैव-सूचना विज्ञान दृष्टिकोण विकसित किया ताकि एकजुटता प्रोटीन संघटन प्रवृत्ति के विभिन्न प्रजातियों को मनमानी रूप से विचार किया जा सके, जो गैर-प्राकृतिक प्रजातियों को शामिल करने के लिए पर्याप्त था। विशेष रूप से, संघटन-प्रवृत्ति घनिष्ठ प्रजातियों के गठन के किनेटिक पथ में एक संरचनात्मक विशिष्ट प्रमुख बाइंडिंग मोड होने का पता चला, जिससे किनेटिक प्रक्रिया मूल प्रोटीन संघटन के समान हो गई। इस तरीके से प्राप्त किनेटिक पैरामीटर सेट का उपयोग एक जनसंख्या संतुलन मॉडल में किया गया और इंसुलिन की चिकित्सा में परिवर्तन के साथ दो आदेशों की विभिन्नता

वाले समय के अनुसार एकत्रिति न्यूक्लीएशन का समय का सटीक पूर्वानुमान मिला, जबकि एक प्राकृतिक पैरामीटर सेट का प्रशंसा परमीकरण के पूर्वानुमान के लिए स्थायी नहीं मिला। इसके अलावा, यह भौतिक रूप से निर्धारित किनेटिक पैरामीटर सेट ने कई यांत्रिक अंशों की समझ प्रदान की, जैसे कि एकत्रिति न्यूक्लीएशन में गति सीमित करने वाली कदम की पहचान और इंसुलिन के पिघलाने के तापमान के चारों ओर अरेनियस से गैर-अरेनियस एकत्रिति किनेटिक्स की ओट देने के लिए एक मात्रावादी पैरामीटर सेट के बदलाव के लिए क्वांटिटेटिव व्याख्या।

इंसुलिन स्व-संघटन में गैर-प्राकृतिक एकत्रिति मार्ग के साथ संरचनात्मक और इंटरएक्शन विशिष्टता स्थापित करने ने हमें इस्त्राकृतिक संघटन न्यूक्लीएशन के प्रारंभिक चरण में शामिल प्रोटीन-प्रोटीन इंटरएक्शन (PPIs) को निरोधित करने के लिए उच्च-विशिष्टता वाले लिगेण्ड्स की पहचान के लिए संरचना-आधारित, कंप्यूटर सहायित विधियों का उपयोग करने की संभावना दी। हम एफडीए अनैक्टिव्स डेटाबेस से वो नेतृत्वक अणुओं की पहचान करने के लिए एक मल्टी-स्तर वर्चुअल स्क्रीनिंग प्रोटोकॉल का उपयोग करके इंसुलिन एकत्रिति को निरोधित कर सकने वाले डिज़ाइन किया। पैराम्यूफोर फीचर्स के द्वारा तीन छोटे पेप्टाइड्स और एक आर्गेनिक फ्लोरोजेन की प्रमुख फार्माकोफोर फीचर्स के आधार पर इनैक्टिव्स डेटाबेस की पहली स्तर की स्क्रीनिंग पर आधारित थी, जिनमें इंसुलिन की एकत्रिति न्यूक्लीएशन को बाधित करने के लिए बाइंडिंग करते समय एकत्रिति-प्रवृत्ति संरचनाओं के साथ बाइंड करने के रूप में दिखाया गया था। फिर हमने इंसुलिन के एकत्रिति-प्रवृत्ति संरचनाओं पर विभिन्न लो-ऊर्जा लिगेण्ड संरचनाओं का एन्सेम्बल डॉकिंग किया, और मॉलेक्युलर डायनेमिक्स सिमुलेशन और बाइंडिंग अफीनिटी गणनाएँ करने के बाद एकत्रिति न्यूक्लीएशन को निरोधित करने के लिए उपयोग किए जाने वाले नेतृत्वक अणु संरचनाओं पर विभिन्न लो-ऊर्जा लिगेण्ड संरचनाओं के डॉक करने के रूप में एक सबसेट की पहचान के साथ एक सबसेट की पहचान करने के लिए मॉलेक्युलर डायनेमिक्स सिमुलेशन और बाइंडिंग अफीनिटी गणनाएँ की। फिर हमने प्रोटीन-लिगेण्ड इंटरएक्शन विवरण और प्रोटीन ओलिगोमरिजेशन के दौरान संरचनात्मक पुनर्व्यवस्थिति प्रभाव को शामिल करने के लिए हमारे डिज़ाइन किए गए एग्रीगेशन न्यूक्लीएशन किनेटिक्स मॉडल का अप्लाई किया ताकि प्रभावी एकत्रिति निरोधन के लिए लिगेण्ड गति का निर्धारण किया जा सके। इन चयनित लिगेण्ड्स (लीड कॉम्पाउण्ड्स) के एकत्रिति निरोधन पर इसके प्रभाव की खोज बड़े पैमाने पर की गई द्वारा, इंसुलिन को विभिन्न गतिसंघटन-प्रवृत्ति जलीय बफर स्थितियों (कम pH, उच्च तापमान) में विभिन्न गतिसंघटन की ताप संरेखण का उपयोग करके किया गया, यहां, आकार विच्छेदन क्रोमेटोग्राफी का उपयोग किया गया और सर्कुलर डिफ्रैक्शन स्पेक्ट्रोस्कोपी का उपयोग किया गया था ताकि गतिसंघटन पैरामीटरों का प्रभाव लिगेण्ड के साथ को-इन्क्यूबेशन के बाद मूल संरचना सामग्री पर कितना होता है। विशेष रूप से, दो लीड कॉम्पाउण्ड्स (राइबोफ्लेविन और एस्पारटेम) जब इंसुलिन: प्रोटीन मौलर गति में 2:1 लिगेण्ड के अनुपात पर प्रयुक्त किए जाते हैं, तो इंसुलिन की एकत्रिति को प्रमुख रूप से विलंबित करते हैं।

दूसरी बायोथेराप्यूटिक जिस पर ध्यान दिया जा रहा है, रिटक्सीमैब, एक आईजीजी1 मोनोक्लोनल एंटीबाडी है जो क्रॉनिक लिम्फोसिटिक लीकीमिया का इलाज करने में प्रयुक्त होता है। लीनियर प्रतिक्रिया द्वारा प्रेरित मॉलेक्युलर डायनेमिक्स सिमुलेशन, धीमे सामान्य मोड के साथ संरचनात्मक संवाद को गति देने के रूप में, कॉन्फिगरेशन स्पेस को प्रभावी और बड़े पैमाने पर अन्वेषित करने के लिए उपयोग किए गए थे, और इस तरह, रिटक्सीमैब के आम्लीय स्थितियों में आंशिक फोल्डेड एग्रीगेशन-प्रवृत्ति संरचनाओं की पहचान की। एक स्थानिक बायोइन्फोर्मेटिक प्रोटोकॉल, जिसमें अमीनो अम्लीयता, औद्योगिक पहुंची हुई सतह क्षेत्र, β -प्रवृत्ति, और प्रोटीन तृतीयक संरचना (3डी संरचना) की जानकारी को एक साथ मिलाने का उपयोग उन आंशिक-फोल्डेड रिटक्सीमैब संरचनाओं में एग्रीगेशन हॉटस्पॉट्स की संभावित पहचान के लिए किया गया था। एक पूरी तरह रिजिड बांडी डॉकिंग के साथ आपर के रूप में आगंतुक को लक्ष्य के रूप में लेकर और फिर संरचनात्मक पुनर्व्यवस्थिति के लिए आपकल्पित गैर-नैटिव डाइमेरिक संरचनाओं की तैयारी के एक मल्टी-स्टेप दृष्टिकोण में द्विमिक्षासन किया गया।

इन संरचनाओं पर MMPBSA गणनाएँ की गईं ताकि गैर-नैटिव बाइंडिंग स्वतंत्र ऊर्जा के योगदान का मूल्यांकन किया जा सके। इन गणनाओं से प्राप्त रैंकिंग के आधार पर, हमने रिटक्सीमैब के एग्रीगेशन को ड्राइव करने में महत्वपूर्ण एपीआर्स की पहचान की (एफसी में छः, फैब में दो, और हिंज में एक)। एफसी क्षेत्र में पहचाने गए एपीआर्स की स्पष्ट स्थानिकता थी, जो आईजीजी1 एंटीबॉडियों के विभिन्न pH शर्तों के तहत डिफरेंशियल स्कैनिंग कैलोरिमेट्री (डीएससी) अन्वेषणों के परिणामों से मेल खाती है। इसके अलावा, फैब क्षेत्र में पाए गए एपीआर्स का पता चला कि वे संप्रेषकता-निर्धारण क्षेत्र (सीडीआर) लूप के भीतर बसे हुए थे, जिनके अंकों में एंटीजन बाइंडिंग में शामिल रहने वाले अमीनो अम्ल शामिल थे। ये फिंडिंग्स रिटक्सीमैब के एग्रीगेशन के पीछे के मेकेनिज़्म के अंदर की विचारों को प्रदान करती हैं और इसके जीव-फार्मास्यूटिकल विकास में संभावित प्रभावों की देन करती हैं।

Table of Contents

Dedication	
Certificate	i
Acknowledgments	iii
Abstract	v
सार	ix
Table of Contents	xvi
List of Figures	xvii
List of Tables	xxxi
List of Symbols	xxxix
Chapter 1	Introduction 1
	1.1 Biotherapeutics – Revolutionizing Medicine 1
	1.2 Aggregation Problem in Biotherapeutics 5
	1.3 Role of Protein Heterogeneity 6
	1.4 Need for Studying Aggregation Kinetics 9
	1.5 Present Work 11
Chapter 2	Literature Survey 17
	2.1 Aggregation mechanism – broad overview 18
	2.2 Prediction of Aggregation Kinetic Rates 19

	2.3 Nature of aggregation prone intermediates	25
	2.4 Virtual Screening	29
	2.5 Effect of Mutation in Insulin	31
	2.6 APR identification in monoclonal antibodies	33
Chapter 3	A Multiscale Model for Quantitative Prediction of Insulin Ag- gregation Nucleation Kinetics	37
	3.1 Introduction	38
	3.2 Theory	38
	3.3 Results and discussion	40
	3.3.1 Reduced Kinetic Model	40
	3.3.2 Diffusional Association Rates	42
	3.3.3 Structure of <i>N.A</i> dimer(s) in optimal binding mode.	44
	3.3.4 Kinetically Dominant Binding Mode(s)	46
	3.3.5 Role of Dimer Conformational Rearrangement.	52
	3.3.6 Prediction of Effect of Protein and Ligand Con- centration on Aggregation Nucleation.	56
	3.3.7 Temperature Dependence of Aggregation Nu- cleation Kinetics	61
	3.4 Conclusion	67
Chapter 4	Multi-level High-throughput Screening for Discovery of Lig- ands that Inhibit Insulin Aggregation	73
	4.1 Introduction	74
	4.2 Results and discussion	75
	4.2.1 Identification of significant pharmacophore fea- tures	76
	4.2.2 Virtual screening of FDA inactives database	80
	Ligand-based screening (LBS)	80
	Structure-based screening (SBS)	82
	4.2.3 Ligand Kinetics	87

	4.2.4	In vitro characterization of aggregation inhibition	89
	4.3	Conclusion	94
	4.4	Experimental and Computational Methods	95
	4.4.1	Computational Methods	95
		Database Screening	95
		Molecular Dynamics	96
		Binding energy calculation	97
		Protein Ligand Diffusional Association Rate	98
	4.4.2	Experimental Method	99
		Sample preparation	99
		Size exclusion chromatography	101
		Circular dichroism (CD)	101
		Thioflavin T fluorescence assay (ThT)	101
		Reverse-phase high pressure liquid chromatog- raphy (RP-HPLC)	102
Chapter 5		Identification of aggregation hotspots in IgG1mAb, Rituximab	103
	5.1	Introduction	103
	5.2	Results and discussion	105
	5.2.1	3D Structure of Rituximab	105
	5.2.2	Identification and Characterization of APRs	107
	5.2.3	Non-native binding modes of Rituximab	115
	5.2.4	Identification of APRs driving aggregation	117
	5.3	Conclusion	125
	5.4	Computational Methods	127
	5.4.1	Homology Modelling	127
	5.4.2	SAP	128
	5.4.3	β propensity	129
	5.4.4	Docking	130
	5.4.5	Molecular Dynamics	130
	5.4.6	Binding energy calculation	132
	5.4.7	Normal Mode Analysis	133

Chapter 6	Summary	135
Chapter 7	Scope For Future Work	139
References		143
Appendix A	Supplementary Information for Chapter 3	197
	A.1 Section of Appendix	197
	A.2 Nucleation time	207
	A.2.1 Point of Inflection Method	207
	A.2.2 Relative Fibril Concentration Method	208
	A.2.3 Smoluchowski Kernel Method	208
	A.3 Methods	209
	A.3.1 Diffusional Association	209
	Coarse Grained (CG) Molecular Dynamics (MD)	209
	All Atom (AA) MD	215
	A.3.2 Linear Integer Programming	216
	A.3.3 Transition Complex Theory	217
	A.3.4 Binding Energy	219
	A.3.5 Dominant Complex	221
	A.3.6 Ligand Association	222
	A.4 Modified Kinetic Rates	224
	A.5 Temperature dependence of aggregation kinetics	226
Appendix B	Supplementary Information for Chapter 4	229
Appendix C	Supplementary Information for Chapter 5	245
	C.1 Unique APRs in a structure, $\overline{\cos\Theta}^{X,Y}$	256
Publications		259
	1 Journal Articles	259
	2 Conference Proceedings	259

List of Figures

3.1	Reduced kinetic model for aggregation nucleation. Concentration of monomer, C_1 , and dimer, C_2 , (blue, left axis), and of fibril, C_F , (red, right axis) as function of incubation time, t , obtained by solving either the full (· ·) or the reduced kinetic model (— —). For both, empirical set of kinetic parameters obtained[120] at 344 μ m insulin, pH 2, and 333 K were used. The aggregation nucleation time ($\blacktriangle$) is shown on time axis.	43
3.2	Energy landscape for $N.A$ complex. (A) Scatter plot for total number of contacts, N_{cr} in the ensemble of structures obtained on a rotation χ away from the bound complex. (B) Standard deviation of the rotational angle, σ_χ , plotted versus N_C . The data is shown for the kinetically dominant $N.A$ dimer. The total number of contacts in the bound state and in the transition state was 32 and 13, respectively. Plots generated using TransComp [228].	47

3.3	Dominant binding mode of $N.A$ diffusional complex. (A) The dimeric complex in surf and ribbon views. The native protein (N) is red and aggregation-prone species (A) is blue. (B) Interaction map of the $N.A$ binding interface determined and plotted using LigPlot.[230] Residues involved in hydrophobic interactions (lilac eyelashes for A and brick red eyelashes for N , pointing towards respective interaction partner) and hydrogen bonds (green dotted lines with donor-acceptor distance in Å) are mentioned (name and sequence position). Black, red, blue, and yellow circles represent carbon, oxygen, nitrogen, and sulphur atoms, respectively.	50
3.4	Effect of protein (insulin) concentration on aggregation kinetics at 333 K. (A) Temporal evolution of fibril concentration, C_F , obtained using the full kinetic model ($\cdots$), and the reduced kinetic models either excluding ($--$) or including ($---$) dimer conformational rearrangement. (B) Aggregation nucleation time (t_{nu}) as function of initial insulin concentration ($[M_0]$) obtained from kinetic models compared with the corresponding experimental estimate ($(*)$ Lee et al.[124] , $(+)$ Amdursky et al.[251]).	55
3.5	Effect of ligand (BSPOTPE) on aggregation kinetics. (A) Interaction map of the $A.L$ complex obtained from LigPlot (interface boundary: dashed line). Representation convention is same as Figure 3.3. (B) Variation in aggregation nucleation time (t_{nu}) with BSPOTPE concentration (L) at 500 μ m insulin ($[M_0]$) obtained by solving either the full model with empirical parameters from ref. [120] ($-\diamond-$) or the reduced model with parameters determined in this work ($-\circ-$). Ligand inhibition was modeled using equation 3.7 in both cases. Experimental estimate is from Hong et al. [235] ($(*)$). (C) Overlay of kinetically dominant $N.A$ and $A.L$ complexes zoomed-in at the binding interface (N : red (surf view), A : blue (surf view), L : green (bonds view)).	58

3.6	Effect of temperature on aggregation kinetics. (A) Variation of aggregation nucleation time, t_{nu} , with incubation temperature as determined in earlier reports by studying aggregation kinetics using fluorescence spectroscopy [99], [124] (blue bar) and corresponding estimate from our model (red bar) wherein dimer conformational rearrangement rate, k_r , was used as an adjustable parameter. (B) Temperature dependence of k_r . (C) Temperature dependence of overall dimer association rate, k_{nu} , (see equation A.27) on secondary y -axis (red $\circ$) and amyloid fibril growth rate, k_{AFGR} , as reported in Arora et al. [99], on primary y -axis (blue *).	62
4.1	Multi-stage Virtual Screening (VS) Protocol. Number of lead molecules left at end of a given step, n , is shown. FDA approved inactives list (FDA_1) was pruned by removing multiple conformations of same entry (FDA_2). Next, molecules were selected based on 2D similarity to dominant pharmacophores of query molecules (LBS) followed by rigid-body docking and ranking based on single structure binding energy and receptor (species A) active site coverage (SBS). The final step consisted of binding affinity and binding kinetics calculations from converged MD simulation trajectories (MD).	76

-
- 4.2 Determination of significant pharmacophore features. (A) Pharmacophores of BSPOTPE, P_i , were defined using Pharmit [278] (atom groupings of each pharmacophore in inset): two charged sulphonate groups (P_1, P_{12}), four aromatic hydrophobic phenyl rings ($P_4 - P_7$), four aliphatic hydrophobic units (P_2, P_8, P_9, P_{11}), and two R – O – R' polar ether group (P_3, P_{10}). The fractional contribution of each pharmacophore to the binding energy, $p(E_i)$ (equation Appendix B: B.2), was obtained from MMPBSA–IE binding energy calculated from eight MD trajectories of the insulin:BSPOTPE complex. A cut-off value of $p(E_i) = 0.3$ (dashed line, – – –) is used to mark dominant pharmacophores (see text for details). (B) $p(E_i)$ from individual MD trajectories on a 2D representation of BSPTOPE (color scale blue (0) to red (1)). (C) BSPOTPE pharmacophores significant to binding ($p(E_i) \geq 0.3$) are shown in a surface mesh 3D view (white). 77
- 4.3 Application of 2D LBS on the inactives database. Tanimoto coefficient calculated from the overlap in the ECFP4 representation of the the database molecule and the selected significant features of the query, T_{cof} (see Appendix B: equation B.4), is reported for (A) BSPOTPE, (B) Pep_1 , (C) Pep_2 , and (D) Pep_3 . The red curve in each panel represents the best fit normal distribution. The mean (μ) and the cut-off value ($\mu + \sigma$, where σ is the standard deviation) are marked by dashed vertical lines. (E) Overlay of five highest-ranked *A.L* complexes (only one pose taken for each ligand) obtained by global docking of inactives database molecules filtered through 2D LBS on partially-folded conformer *A* of the kinetically dominant $N.A^5$ insulin dimer. Ligands are in licorice representation (αTA (green), Rib (pink), γTc (blue), PPP (purple), δTc (orange); see Appendix B: Table B.7 for full chemical names), and species *A* is in licorice mesh view (red), its aggregation hotspot in the $N.A^5$ dimer in white. . . . 81

4.4	Kinetic model estimate for aggregation nucleation time. Nucleation time, t_{nu} , was calculated using an insulin aggregation kinetics model[94] for a 600 μ m ($= [M_0]$) insulin formulation at pH 2 and 333 K for various ligand concentrations (reported as molar ratio, $[L]:[M_0]$). Data is reported for Rib ($- \circ -$), Bez ($- \diamond -$), Asp ($- \square -$), BeB ($- * -$), and α TA ($- + -$).	87
4.5	Aggregation inhibition with two highest efficacy inactives. Insulin was incubated under amyloidogenic conditions with either riboflavin (Rib) or aspartame (Asp), or with no stabilizer (control). Temporal evolution of retained monomer (determined from SEC) is shown for (A) 100 μ m and (B) 600 μ m insulin formulations. The data is for the control (black *), formulations with Rib at $[L]:[M_0]= 1:1$ (blue +, only at 600 μ m) and 2:1 (blue $\circ$), formulations with Asp at $[L]:[M_0]= 2:1$ (red $\square$). (C) Far-UV CD spectra at the end of incubation for the 600 μ m formulations: control at 0 h (black —) and 48 h (black —), 1:1 (blue —) and 2:1 (blue —) Rib formulations at 48 h, 2:1 (red —) Asp formulation at 48 h.(D) ThT growth curves for the 600 μ m formulations: control (black *), formulations with Rib at $[L]:[M_0]= 1:1$ (blue +) and 2:1 (blue $\circ$), formulation with Asp at $[L]:[M_0]= 2:1$ (red $\square$). Bars represent the standard deviation. Panels A and B (SEC) have data from two technical replicates and Panel C (ThT) has data from four technical replicates.	90

4.6	<i>A.L</i> interface interaction map of Rib and Asp. 2D chemical fingerprint similarity to the parent query molecule for (A) Rib (parent: BSPOTPE, significant pharmacophores circled) and (B) Asp (Parent: Pep ₃ , only significant pharmacophores shown) are shown by lines connecting matching atom pairs. <i>A.L</i> interface interactions involving these specific atoms of <i>L</i> and hotspot residues of conformer <i>A</i> are shown for (C) Rib and (D) Asp. Conformer <i>A</i> is taken from the highest inhibited binding mode: <i>N.A</i> ⁵ for Rib and <i>N.A</i> ¹ for Asp. Hydrophobic interaction pairs are denoted by red brick eyelashes and red dashed lines; hydrogen bonds by green dotted lines (donor-acceptor atom distance is mentioned). Black, red, blue, and yellow circles represent carbon, oxygen, nitrogen, and sulphur atoms, respectively.	91
5.1	A flowchart illustrating a concise overview of the protocol employed in the present study.	104
5.2	3D structure of Rituximab: The structure of Rituximab was obtained through homology modelling using the X-ray structure of the Fab domain (PDB ID – 2OSL) and Fc domain (PDB ID – 1L6X) as template and modelling the hinge region. The obtained structure has been shown in blue ribbon view and hinge region has been highlighted in blue surf view. The X-ray crystal structure of Fab and Fc (red ribbon view) has been overlapped with the modelled structure. All 16 di-sulphide bridges have been shown in green ball and stick view.	105

5.3	Sample structure obtained through MDeNM: APRs of a conformation obtained through MDeNM (St_{NM}^{18} in Table 5.1a) has been compared against base structures from APR#1 [St_B^1 and St_B^2 in Table 5.1a]. The function $f(\overline{\cos \Theta}^{St_{NM}^{18}, St_B^i}) = \min(\max(\overline{\cos \Theta}^{St_{NM}^{18}, St_B^1}), \max(\overline{\cos \Theta}^{St_{NM}^{18}, St_B^2}))$ reports the APRs unique to St_{NM}^{18} when compared with St_B^1 and St_B^2 and has been shown as a bar plot (Panel A). Two APR is unique in the given mode (one each in V_H and C_L) and marked by ▼ over the bar. In panel (B), a 3D snapshot showcases the aforementioned two APRs. The green arrows represent the direction vectors (identified from the corresponding normal mode 18) along which St_{start} was excited. The unique APR exposed as a result of this excitation has been shown in green surf view (one in V_H and one in C_L region)	110
5.4	Final ensemble of partially folded aggregation prone conformations of Rituximab: 3D ribbon view of the final set of ten partially folded near native conformations (shown in green) overlaid on the native structure (shown in red). Secondary structure content has been shown in Appendix C: C.18	113
5.5	Binding affinity of poses obtained from flexible docking: $\overline{B.E}_{dock}$ obtained from Prodigy has been plotted (as ○) in decreasing order for all obtained poses. A six degree polynomial was fit (shown in yellow —) and the first plateau was taken as the cut-off (shown in black —), resulting in a cut off $\overline{B.E}_{dock}$ of $-17 \text{ kcal mol}^{-1}$ and selection of 57 poses (shown in inset).	116
5.6	Energetically dominant APRs in putative aggregation-pathway rituximab dimers: (A) $\overline{\Delta G}_{APR^i}$ of obtained APR clusters (see Method section 5.4.6 for details) in decreasing order of favourable binding affinity with the cut-off show in (—). The top nine APR shortlisted are shown in (B) licorice view coloured w.r.t to residue name on the 2D representation of mAb domains and (C) surf view (red) on native Rituximab molecule with Fab fragment shown in blue ribbon and Fc fragment shown in green ribbons.	118

A.1	Plane Fitting on interface: The plane fitted to the C_α atoms of the interface residues of the native protein, N . The aggregation prone protein A moved along the normal by a distance of Debye length showing its new position A''	198
A.2	Convergence of MD trajectories: $\cos \Theta^{N.A;N.A}$ tracked for every frame during the CG–MD trajectory for all the parent structures is reported showing a stable value for at least 200 ns	199
A.3	Family of structures: The complexes obtained at the end of MD simulations are categorized into families based on their common distinct polar "donor–acceptor" contacts on their interface according to the work of Gabdoulline et al. [42]. Complexes belonging to the same family have two or more common polar contacts and one or zero common polar contacts with the complexes of a different family. (A) Distribution of obtained structures among various families. (B) The heavy atom inter–atomic distance between inter–protein residues need to be less than 0.5 nm to be considered as a contact and the distance between two intra–protein contact needs to be more than 0.6 nm to be considered as a distinct contact.	201
A.4	Identification of dominant complex. (A) The concentration profile of $[NA^i]$ obtained by solving the equations A.14 and A.15 using the parameters listed in Table 2 ,showing the existence of a dominant complex. (B) Interface of the dominant complex showing the hotspot residues for A in blue (I) and N in red (II) in panel	202
A.5	Kinetically favourable binding modes: Binding modes identified in the present study with favourable binding energy ($\Delta G \leq 0$). The complexes have been shown in ribbon and surf view with A shown in blue and N shown in red.	203

A.6	Calculation of k_r by fitting monomer data: Value of k_r in Equation A.27 was obtained by using it as a fitting parameter and fitting to the monomer data of Murray et al [120]. upto the nucleation time. Goodness of fit has been shown in the figure with ($\cdots$) showing Murray et al monomer data without fibrillation and ($-$) showing monomer data in present study upto t_{nu}	204
A.7	Calculating k_{aL} : (A) Interface area of the AL complex during the simulation . (B) $\cos \Theta^{AL;AL}$ calculates the inner product between the A of the final AL complex and A of the complexes formed during the trajectory (C) The standard deviation of the rotational angle (χ), σ_χ , is plotted as a function of total number of contacts, N_c for the $A.L$ complex. The transition state was determined from the distribution $\sigma_\chi(N_c)$. (D) Scatter plot for the total number of contacts, N_c , for the $A.L$ on a given rotation (specified by the rotational angle χ) away from the bound complex. The data is shown for the highest ranked $A.L$ complex where the total number of contacts in the bound state and in the transition state was equal to 17 and 5, respectively.	204
A.8	Effect of ligand on insulin aggregation nucleation time. (A) Nucleation time t_{nu} calculated for three different K_{AL} values: ($- \circ -$) $K_{AL} = 10 \mu m$, ($- + -$) $K_{AL} = 5 \mu m$ and ($- \times -$) $K_{AL} = 1 \mu m$. ($*$) represents data from Hong et al. [235]. (B) Nucleation time calculated using the reduced kinetic model with parameter values for protein folding and various protein-protein association steps taken from Murray et al. [120] and parameters for protein-ligand association-dissociation kinetics were taken from the present study.	205
A.9	Temperature scaling of kinetic parameters Linear scaling of log (rate) against $1/T$ for kinetic parameters K_{A2} , K_{NA} , K_u , k_a , k_r , k_u	206
A.10	Obtaining t_{nu} from sigmoidal curve of fibril concentration, C_F : Concentration profile of C_F for a $[M_0] = 344 \mu m$ with the tangent at the point of inflection shown.	207

A.11	Stability of partially folded intermediates, PF_i , with MARTINI CG force-field: RMSD of the C_α backbone atoms of the native (yellow), PF_1 (pink), PF_2 (red) and PF_3 (green) after running a simulation with MARTINI force field with ElNeDyne elastic restraints on the secondary structures of the complex.	210
A.12	Scaled vs Unscaled MARTINI: As a graphical representation, we report two trajectories obtained through (A) Scaled parameters (B) Unscaled parameters. Scaled parameters trajectory shows sufficient variation in the interface area of $N.A$ as compared to original parameters trajectory where the complex is formed almost instantaneously with an almost constant interface area.	213
B.1	Trajectories in which pharmacophore features (a) P_1 and (b) P_{12} are buried ($RASA_x \leq 60\%$) or exposed ($RASA_x \geq 60\%$). Frames in which P_i is buried was given a flag value of 1 or 0 for exposed for the full length of trajectory. Cells in red represent buried (trajectories in which more than 50 % of frames are 1) and in blue represent exposed for a given P_i	230
B.2	Significant residues of peptide set The peptide set (A) Pep_1 , (B) Pep_2 and (C) Pep_3 shown in red ribbon view (surface covered in mesh view) with the hexameric Arg hydrophilic tail shown in blue ribbon view. The significant residues have been shown in white licorice view.	230
B.3	Structure Based Screening: A from $N.A^i$ and its aggregation prone hotspot shown in red and white surface mesh view respectively. The top pose for various ligand (α TA (green), Rib (pink), γ Tc (blue), PPP (purple), δ Tc (orange); see Table B.7 for full chemical names) has been shown in licorice view. (A) $N.A^1$ (B) $N.A^2$ (C) $N.A^3$ (D) $N.A^4$	231
B.4	RP-HPLC chromatogram of excipients incubated in fibrillation buffer (0.01 N HCl and 0.1 N NaCl, pH 3.0) at 40 °C A) Riboflavin (0 h and 48 h); B) aspartame (0 h and 48 h); C) Benzoin (0 h and 48 h)	237

B.5	Far-UV CD spectra at the end of incubation for the 600 μm formulations: (A) control at 0 h (B) control at 48 h (C) 2:1 Asp formulation at 48 h (D) 1:1 Rib formulation at 48 h (E) 2:1 Rib formulation at 48 h. Bars represent the signal/noise ratio calculated from three repeat scans (accumulations) on one sample.	238
B.6	Monomer retention % of insulin obtained from SEC (for $[M_0] = 600 \mu\text{m}$) for control (*) and in presence of ligand ($\circ$) (A) BeB (B) Bez (C) αTA (D) DMSO	239
B.7	Monomer retention % of insulin obtained from SEC in presence of (a) designed Peptides of Mishra et al[62]. in 600 μm insulin formulation with peptides (Pep_1 (*) Pep_2 ($\circ$) Pep_3 (+)) present in $[L]:[M_0]$ 2:1, pH 3, 340 K; control ($\square$) (b) BSPOTPE (*) at $[L]:[M_0]$ 1:1 in 100 μm insulin formulation (conditions same as for Riboflavin and Aspartame in the present work).	239
B.8	SEC chromatograms of (A) 600 μm insulin at 0 h (Control, Rib (1:1, 1:2), Asp (1:1, 1:2)); (B) 600 μm insulin at 48 h (Control, Rib (1:1, 1:2), Asp (1:1, 1:2)); (C) 100 μM insulin at 0 h (Control, Rb (1:1, 1:2), Asp (1:1, 1:2)); (D) 100 μm insulin at 48 h (Control, Rib (1:1, 1:2), Asp (1:1, 1:2)); (Inset) calibration curve of amount of insulin vs area under the curve. $[M] = \text{AUC}/(1.3382 \times \text{volume of sample})$; $[M]\%$ retention = $\frac{M_t}{M_0} \times 100$; $[M_t]$ = Concentration of insulin at time t; $[M_0]$ = Concentration of insulin at time 0.	241
B.9	Temporal profile of $[M]$: Temporal profile of $[M]$ obtained from our multi-scale predictive model plotted along experimental SEC data at two different $[M_0]$ i.e (A) 100 μm and (B) 600 μm . $[M]$ obtained from computational model for Rib-insulin solution at 2:1 $[L]:[M_0]$ (blue-), Rib-insulin solution at 1:1 $[L]:[M_0]$ (blue--), Asp-insulin solution at 2:1 $[L]:[M_0]$ (red -). SEC data $[M]$ of Rib-insulin solution at 2:1 $[L]:[M_0]$ (blue $\circ$), Rib-insulin solution at 1:1 $[L]:[M_0]$ (blue +) and Asp-insulin solution at 2:1 $[L]:[M_0]$ (red $\square$)	242

B.10 Temporal profile of [M]: (A) Interface area of BSPOTPE $A.L$ complex during the simulation. (B) $\cos \Theta^{AL;AL}$ calculates the inner product between the A of the final AL complex and A of the complexes formed during the trajectory	243
C.1 Charged Residues along with their sequence number for Rituximab at pH = 3.5, 310 K	246
C.2 Backbone C_α RMSD of (a)Fab and (b)Fc fragment tracked over a 30 ns MD simulation trajectory in native formulation conditions (pH 6.5, T = 277 K)	246
C.3 TM-score reported for the trajectory under formulation conditions (pH = 6.5, T = 277 K) calculated w.r.t to the starting structure of the trajectory.	247
C.4 Residue wise RMSF, calculated from trajectory under formulation condition, has been reported. Flexible region has been marked with vertical lines, where A: CDR H1, B: CDR H2, C: CDR H3, D: Hinge, E: A: CDR L1, F: CDR L3.	247
C.5 Secondary structure content of the trajectory obtained from formulation condition, with % Coil (hydrogen bonded turn, bend, loops and irregular elements) shown in black, % Strand (Residue in isolated beta-bridge, Extended strand, participates in beta ladder) shown in red and % Helix (α -helix, 3-helix (3/10 helix), 5 helix (π -helix)) shown in green.	247
C.6 C_α RMSD of generated modes against the initial structure.	248
C.7 Backbone C_α RMSD of (a)Fab and (b)Fc fragment tracked over a 30 ns MD simulation trajectory in non-native amyloidogenic conditions (pH 3.5, T = 310 K)	248
C.8 Cluster population obtained on applying Gromos clustering algorithm on non-native MD trajectory with a C_α cut-off of 0.2 nm	248
C.9 C_α RMSD of generated modes against the initial structure.	249

C.10	Obtained frequency (in ps^{-1}) of the modes calculated from the modelled elastic network. The time-period has been shown in the inset.	249
C.11	Cumulative contribution of slowest modes to the overall conformational variance. The mode index has been shown in decreasing order of their frequency in the x-axis	249
C.12	(A) Correlation values at various cut off radius, with highest correlation observed at $R_c = 11 \text{ \AA}$ (B) RMSF values obtained from low pH simulation (shown in blue) and excited normal mode simulation (shown in red)	250
C.13	Sample structure obtained through linear combination of shortlisted normal modes:Excitation along linear combination of selected normal modes provided new structures containing APRs of each mode in a single structure. An example of such sampled structure, $\text{St}_{\text{LC}}^{8702}$ containing all the APRs of shortlisted modes St_{B}^4 and St_{B}^5 . APRs in each structure have been shown in black stripes in their respective domains. Unique APRs to the structure have been shown in red stripes. All APRs (including unique APRs) of St_{B}^4 and St_{B}^5 was present in $\text{St}_{\text{LC}}^{8702}$.	250
C.14	Backbone C_α RMSD of the obtained structures through linear combination (St_{LC}^i) calculated against the starting structure (St_{start}) ..	250
C.15	Backbone C_α RMSD of the obtained structures through linear combination (St_{LC}^i) calculated against the native structure (St_{EM}) ..	251
C.16	Secondary structure content of the ensemble obtained from linear combination excitation of slow modes, with % Coil (hydrogen bonded turn, bend, loops and irregular elements) shown in green, % Strand (Residue in isolated beta-bridge, Extended strand, participates in beta ladder) shown in blue and % Helix (α -helix, 3-helix (3/10 helix), 5 helix (π -helix)) shown in red. Secondary structure content of St_{EM} have been shown in black line.	251

C.17	Categorisation of interface contacts: All distinct interface contacts were identified using the protocol applied in other computational studies on protein–protein association[62], [94], [95], [318], [319]. All heavy atoms with inter–atomic distance $\leq 5 \text{ \AA}$ were listed and all contacts with intra–atomic distance $\leq 6 \text{ \AA}$ were categorised as dependent contacts to obtain a final list of independent interface contacts for all 57 poses. The interface contacts were categorised on APR type, and fraction of total contacts belonging to APR#1–APR#1, APR#2–APR#2 and APR#1–APR#2 have been shown as bar plot. The red and green bar shows values for docked complexes and docked complexes after MD refinement respectively	252
C.18	Secondary structure content of the final ensemble of partially folded conformations with Coil (hydrogen bonded turn, bend, loops and irregular elements) shown in red, Strand (Residue in isolated beta–bridge, Extended strand, participates in beta ladder) shown in blue and Helix (α –helix, 3–helix (3/10 helix), 5 helix (π –helix)) shown in yellow.	253
C.19	Stability of trajectories were observed by clustering them using FCC algorithm and reporting the fraction of trajectories present in the most populated cluster. Various values of FCC was used as cut–off {0.10, 0.25, 0.40, 0.75, 0.80, 0.85} and fraction of trajectory present in top cluster was reported for all the poses as a bar plot.	255
C.20	Plot of first derivative (dy/dx) calculated for the 8 degree polynomial $((-3.295 \times 10^{-13})x^8 + (1.492 \times 10^{-10})x^7 - (2.816 \times 10^{-8})x^6 + (2.862 \times 10^{-6})x^5 - (1.692 \times 10^{-4})x^4 + (5.832 \times 10^{-3})x^3 - (1.131 \times 10^{-1})x^2 + 1.176 \times 10^0x^1 - (2.403 \times 10^1))$ fit to $B.E_{\text{dock}}$ of first 100 poses in Figure 5.5B.	256
C.21	Flowchart explaining the protocol for calculation of $\overline{\cos}^{X,Y}\Theta$	257

List of Tables

2.1	List of commonly used protein aggregation models in literature	19
2.2	Coarse grained models with residue level resolution.	20
2.3	A list of models available in the literature with the method used and external environmental factors studied.	24
2.4	Experimental studies reporting aggregation proceeding through a partially folded intermediate with the name of the protein, conditions and experimental used	26
2.5	Enhanced sampling techniques to explore protein conformational landscape	28
3.1	Aggregation kinetics schemes. Scheme A: early events of aggregation involving inter-conversion between native protein (N) and an aggregation-prone monomeric species (A_1 , also referred to as A for ease of notation) (rates: k_f, k_u) and sequential addition of A to various species to form higher order oligomers (A_i) up to a critical nucleus of size n_c^* (rates: k_{nu}, k_{-nu}). Scheme B: same as Scheme A, but protein association step explicitly represents diffusional association (rates: k_a, k_d) and conformational rearrangement (rates: k_r, k_{-r}). Scheme C: addition of ligand to aggregation-prone species (rates: k_{aL}, k_{dL}). Scheme D: formation of higher-order fibrillar species and fibril fragmentation (rates: $k_{fib}, k_{fib-}, k_{frag}$). For all schemes, $1 \leq i < n_c^*$. All species of size $\geq n_c^*$ are labelled F.	39

3.2	Kinetic and thermodynamic parameters for eight binding modes of $N.A$ dimer. The association rate of $N.A$ dimer, $k_{a,}$ and the electrostatic interaction energy of the corresponding transition state complex, $\Delta G_{N.A}^{*(elec)}$, were determined using the transition complex theory. The binding free energy of the $N.A$ dimer, $\Delta G_{N.A}^o$, was calculated using the MM–PBSA method on a 20 ns MD trajectory. It also includes the contribution from protein configurational entropy change estimated using an atom specific weighted surface area (WSAS) method [229].	49
3.3	Kinetic parameters in insulin aggregation nucleation. Column A : Present work; Column B : Full set of parameters determined empirically by fitting to aggregation growth data [120]; Column C : direct experimental measurements for individual elementary steps for insulin or other globular proteins.	52
4.1	Lead candidates from the inactives database after SBS. Top nine molecules (abbreviation in column P, see Appendix B: Table B.7 for full chemical name) selected against each A conformer from five $N.A^i$ dimers, identified on basis of ligand–binding affinity from Vina (G_{dock}^{vina} in column R), are listed. The molecules are arranged in increasing order of MMPBSA–IE binding energies calculated from MD trajectories ($G_{sim,}$ in Column Q). The molecules with limited aqueous solubility are shown in red and those taken forward for in vitro characterization are in green (see text for details). All binding energies are reported in kcal mol^{-1}	83

4.2	Insulin aggregation dependence on ligand concentration. Insulin was incubated for 48 h under amyloidogenic conditions (pH 2, 313 K, 0.1 N NaCl) in presence or absence (control) of screened FDA inactives. Insulin monomer retention at end of incubation (determined using SEC) is reported for 100 μ m and 600 μ m formulations at various lig- and:protein molar concentration ratios, [L]:[M ₀]. For comparison, the inhibitory effect of the query molecules (peptides and BSPOTPE) is shown in Appendix B: Figure B.7 and Appendix B: Table B.10. For a 600 μ m insulin formulation (T = 333 K, pH 1.6), the peptides were able to inhibit insulin aggregation for 72 hrs at a [L]:[M ₀] 2:1 whereas BSPOTPE inhibited 500 μ m insulin for upto a week for [L]:[M ₀] 1:1[61], [62]	89
5.1	Distribution of sampled APRs on various domains of Rituximab: The distribution of aggregation prone regions (APRs) in the shortlisted configurations were analyzed on a per-domain basis. The results were presented in the form of (m, n), where m represents the total number of APRs, and n represents the number of unique APRs in the particular domain. The values of (m, n) with n > 0 were shown in bold. Distribution of APR#1 has been shown in (a) and APR#2 in (b).109	
5.2	Distribution of APR Across Antibody Domains: Categorization of APRs according to their specific antibody domain location. 'Nos of APR' indicates the total count of APRs identified within each domain. The 'Buried in Native' and 'Exposed in Native' columns detail the number of these APRs that are buried or exposed, respectively, in the antibody's native state	115

-
- 5.3 Summary of APR Characteristics Across Different Antibody Domains: Enumeration of dominant APRs, detailing their location within the antibody domain (Domain), their exposure status in the native conformation (Status in Native), and the percentage of secondary structure retention after unfolding (Sec. Structure Retention). Solvent accessible surface area (SASA) measurements for each APR are provided both for the native state (SASA in Native) and for the partially folded (PF) state (SASA in PF). The 'APR Exposed Count in PF' column quantifies the occurrences of the APR being exposed across identified 12 partially folded instances. The 'Docked Interface Count' reflects the prevalence of APRs at the interface in the 57 putative non-native docked complexes. The '% of APR Validated' column, along with its 'Activity' and 'Aggregation' sub-columns, indicates the proportion of the APR sequence previously validated for functional activity and importance in aggregation, respectively. 119
- 5.4 Sequences Involved in Shortlisted APRs: Amino acid sequence of shortlisted APRs with their domain wise location and their type (APR#1 or APR#2). The secondary structure of each sequence has been reported in column Z in the form of (P,Q) where P is the secondary structure in the native structure and Q in the aggregation prone conformation (C: hydrogen bonded turn, bend, loops and irregular elements; E: Residue in isolated beta-bridge, extended strand, participates in beta ladder; H: α -helix, 3-helix (3/10 helix), 5 helix (π -helix)). Energetic contribution of each amino acid to its $\overline{\Delta G}_{APR^i}$ has been reported in column Y. The sequences have been colour coded with hydrophobic residues shown in green, polar residues in red, charged residues in blue and aromatic residues in black. APRs identified in previous experimental studies on aggregation/functional activity has been denoted in bold, with experimentally validated residues denoted in italics. 120

A.1	Smoluchowski kernel rates: Higher order oligomer association rates k_{nu}^{ij} (rate of association of A_i and A_j) in terms of k_{nu}^{11} as $k_{nu}^{ij} = \alpha \times k_{nu}^{11}$	197
A.2	Multi-scale Modelling Schematic: A summarized table of the objective of each step in the multi-scale model and the computational technique used to achieve it	198
A.3	Interface area of the structures obtained at the end of CG run and after extending it in its back-mapped AA form (in nm ²). The structures in bold are the parent structures.	199
A.4	Inner product, $\cos \Theta^{N.A^i;N.A^j}$ of the structures obtained at the end of CG run (Column A) and after extending it in its back-mapped AA form against its parent initial docked structure (Column B). The highest $\cos \Theta^{N.A^i;N.A^j}$ of each diffusional complex with any of the initial docked structure is reported (Column C). The structures in bold are the top ensemble structures obtained after LIP.	200
A.5	The binding interface area of top 16 docked structures is listed. The highest $\cos \Theta^{N.A^i;N.A^j}$ of docked structure against a set of diffusional complex obtained from its own set of simulation is reported in Column A. The highest $\cos \Theta^{N.A^i;N.A^j}$ of docked structure against the whole set of all diffusional complexes is reported in Column B.	200
A.6	Aggregation Hotspot in A^i : Residues of A^i involved in binding in $N.A^i$ with their energetic contributions (ΔH and $-T\Delta S$ in kcal mol ⁻¹) to the overall binding energy ΔG . Residue X ^{Ai} represents residue X with residue number i belonging to chain A.	203
A.7	κ_n at various initial monomer concentration, $[M_0]$	205
A.8	The temperature dependence of overall dimer dissociation constant (K_{D,A_2}) and the unfolding equilibrium constant (K_u), was determined using the Van't Hoff equation, wherein the standard enthalpy change (ΔH^0) and the standard entropy change (ΔS^0) of the reaction are taken to be constant. These values were taken respectively from Strazza et al. [231] (for K_{D,A_2}) and Choi et al.[253] (for K_u).	206

A.9	Sensitivity analysis for all kinetic parameters in our model between the temperature range of 315.5 K – 373 K by varying the values of each parameter sequentially at the mentioned temperature and measuring t_{nu} . We report τ where $\tau = \frac{t_{nu}^{315.5\text{ K}} - t_{nu}^{373\text{ K}}}{(t_{nu}^{315.5\text{ K}} - t_{nu}^{373\text{ K}})_{exp}}$. The parameters with τ values greater than one were considered significant in our model and Relative Root Mean Square Error (Equation A.28) for those parameters were reported. Opposite signs for $K_{D,N.A}$ and K_u, k_r signifies they are oppositely correlated in the model.	206
A.10	As MARTINI overestimates the short range interactions, the forcefield parameters were scaled down. The scaling down parameter α was determined by running MD simulation for 100 ns on each of the 16 top ranked complexes obtained from Avinash et al[62]. and calculating $S_{\%}$, the percentage of total simulations in which the starting complex remained stable for at least 50 ns. The scaled down well-depth, ϵ , in Lennard Jones 12–6 potential function for a scaling factor of α is given by $\epsilon = \epsilon_{repulsive} + \alpha \times (\epsilon_{original} - \epsilon_{repulsive})$ where, $\epsilon_{repulsive}$ is 2.0 kJ/mol, and $\epsilon_{original}$ is the original MARTINI well-depth used for the interaction of interest	212
A.11	Unfolding rate, k_u , at 333 K and its activation energy, E_a , have been taken from Singh et al.[93] where metadynamics simulations were used to obtain kinetics and thermodynamics of insulin unfolding. The temperature dependence of diffusional association rate, k_a , as calculated from equation A.10, was determined by taking the electrostatic interaction energy of the transient state complex, $\Delta G_{N.A}^{*(elec)}$, as a constant and modeling temperature dependence of the basal association rate by the Stokes–Einstein equation($\frac{k_0^{T_1}}{k_0^{T_2}} = \frac{T_1\mu_2}{T_2\mu_1}$ where μ is the dynamic viscosity).	226

A.12	For temperature dependence of the diffusional dimer dissociation constant, $K_{D,N.A}$, equation C.1 was used with ΔE_{MM} ($-49.56 \text{ kcal mol}^{-1}$) and configurational entropy ($\Delta S = -111.15 \text{ cal mol}^{-1}$) taken independent of temperature, while the temperature dependence of apolar surface tension taken from Eötvös rule [254], $\gamma V^{\frac{2}{3}} = k(T_c - T)$ where V is molar volume (assumed constant in present study), k is a universal constant ($2.1 \times 10^{-7} \text{ JK}^{-1}\text{mol}^{2/3}$) and T_c is the critical temperature (647 K in present study).	227
B.1	Calculated $p(E_i)$ for all pharmacophore features, P_i , of BSPOTPE in each trajectory	229
B.2	$p(E_i)$ for each residue of the synthetic peptides with significant residues ($p(E_i) \geq 0.30$) shown in bold.	231
B.3	Interface similarity, $\cos \Theta^{A^m, A^n}$, of A^i among non-native binding modes $N.A^i$, ($\cos \Theta^{A^m, A^n} = 0$ (no overlap), 1 (full overlap)):	231
B.4	Backbone C_α RMSD of all A^i identified in $N.A^i$ against each other, $C_\alpha^{A^m, A^n}$.	231
B.5	Interface similarity ($\cos \Theta^{N.A^i; A.L}$) distribution obtained upon docking 432 screened ligands from ligand based screening (all top nine complex per ligand considered) to the identified binding modes. Column A: Distribution for all 432 screened ligands. Column B: Distribution for top 50 (ranking based on $\Delta G_{\text{dock}}^{\text{vina}}$) ligands.	232
B.6	Interface similarity ($\cos \Theta^{N.A^i; A.L}$) for final list of ligands (see Table 2)). Column A: Maximum and minimum $\cos \Theta^{N.A^i; A.L}$ of the top nine complexes of each ligand against its parent binding mode. Column B: $\cos \Theta^{N.A^i; A.L}$ of the best pose (most negative $G_{\text{dock}}^{\text{vina}}$) of the ligand against its parent binding mode.	232
B.7	Acronyms for shortlisted FDA inactive hit molecules. Molecules with solubility $\leq \frac{600}{10} \mu\text{M}$ are shown in red.	232
B.8	Interface area of $A.L$ for the hits obtained from each binding mode. Column X and Y represent the area of $A.L$ in docked complex and after molecular refinement respectively in nm^2	234

B.9	Diffusion–Limited Association Rates: k_a and k_{aL} (obtained from SDA7) for the dominant binding mode, $N.A^5$ and $A.L_{X,N.A^i}$ respectively where X is the inhibitor under study and $N.A^i$ is the binding mode from which X is obtained. Dissociation constant $K_{D,N.A}(= \frac{k_d}{k_a})$ and $K_{D,A.L}(= \frac{k_{dL}}{k_{aL}})$ were obtained using MM/PBSA–IE through g_mmpbsa[317]	236
B.10	Monomer retention % of insulin obtained from the work of Hong et al[61] (Data reported for an incubation time of 12 h, $[M_0] = 500 \mu\text{M}$, pH = 1.6, and T = 338 K, no shaking or stirring.)	240
B.11	Deconvolution of the CD data for calculating secondary structure content (at the end of 48 h incubation for $[M_0] = 600 \mu\text{M}$ pH 2 (0.01 N HCl), 313 K, 0.1 N NaCl with constant shaking at 300 rpm.) was performed using Bestsel web server[308].	240
C.1	Amino acid sequence of Rituximab in FFASTA format	245
C.2	Two pairs of conformations, one from APR#1 and the other from APR#2, were specifically chosen based on their high APR cluster counts. These conformations were subjected to docking with the APRs as the target. Consequently, two distinct sets of docked complexes were generated: set 1 comprising 1000 poses, and set 2 comprising 250 poses. The fraction of top structures in set 2, exhibiting a similar interface (interface RMSD, iRMSD $\leq 7 \text{ \AA}$) to those in set 1, was determined for both complexes derived from APR#1 and APR#2. The comparison was performed for the top 10, 15, and 30 structures.	255

List of Symbols

χ Rotational angle

κ_n Relative fibril concentration at t_{nu}

$p(E_i)$ Boltzmann averaged energetic contribution of a pharmacophore feature to binding energy

T_{cof} Tanimoto coefficient

A Aggregation Prone Conformation

$\cos\Theta$ Interface similarity index

D Denatured conformation

f Fractional monomer conversion

k_{aL}, k_{dL} Ligand association/dissociation rate

k_a, k_d Non-native association/dissociation rate

k_{fib}, k_{-fib} Fibril growth/dissociation rate

k_{nu}, k_{-nu} Nucleus growth and dissociation rate

k_r, k_{-r} Conformation re-arrangement rate

k_u, k_f Unfolding, Folding rate

N Native conformation

N_c Total number of inter-atomic contacts

N_c^*	Total number of inter-atomic contacts in the transition state
n_c^*	Critical nucleus size
P_i	Pharmacophore feature
T_m	Protein melting temperature
t_{nu}	Nucleation time