Mapping the Structural Drivers of Insulin Recognition and Specificity Using Molecular Dynamics and Free Energy Calculations

Mohan Maruthi Sena*[1], Ramakrishnan. C*[1], M. Michael Gromiha[2], Monalisa Chatterji[1], Anand Khedkar[1] & Anirudh Ranganathan[1#]

1. Sekkei Bio Pvt. Ltd., Third Floor, Bio-incubator, IITM-Research Park, Kanagam Road, Chennai-600113, India.
2. Protein Bioinformatics Lab, Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai-600036, India.

*Contributed equally

Abstract:

Insulin has been the cornerstone of diabetes treatment since its discovery over a century ago. Despite this, several aspects of insulin engagement with its target and off-target receptors remain unknown, thus limiting the use of structure-guided design of novel analogs. A recent spurt in structural information for the insulin and insulin-like growth factor 1 receptors (IR and IGF-1R), were leveraged in this study to gain a deeper, molecular-level understanding of ligand recognition at these targets. Molecular dynamics (MD) and free energy perturbation (FEP) were utilized to map the drivers of potency and specificity of several insulin variants including disease-linked mutations. The remarkable accuracy of MD-FEP in capturing functional trends (>80% of cases) across various insulin analogs underscored the method's potential. The ability to deconvolute observations into receptor contributions, conformational perturbations and (or) solvent-network changes lays the foundations for its use in predictive design of future insulins and other therapeutic peptides.

Introduction

Insulin is an anabolic peptide hormone that plays a crucial role in glucose homeostasis.[1–3] Mature bioactive human insulin has been structurally well-characterized and is a 51 amino acid peptide consisting of two chains A and B held together by two inter-chain disulfide bridges and further stabilised by an intra-chain-A cysteine-bridge.[4–8] Insulin mediates a host of important metabolic functions through its engagement with the insulin receptor (IR).[1,4,9,10] The IR belongs to the receptor tyrosine kinase (RTK) family of receptors that include type 1 and 2 insulin like growth factor receptors (IGF1- and IGF-2R) and the orphan insulin receptor-related receptor.[4,5,9] The RTKs share a common architecture consisting of an extracellular domain (ECD), a single-spanning transmembrane region, and an intracellular auto-catalytic kinase domain.[5] IRs and IGFRs belong to Class-II (RTK) receptors which form functional homodimers stabilized through numerous cysteine bridges.[4,5,7] These homodimers undergo conformational rearrangements upon ligand binding that results in auto-phosphorylation of specific tyrosine residues intracellularly and downstream signaling.[4,10] Key physiological functions mediated by IR activation include glucose uptake, glycogen synthesis in skeletal muscles, diminished glucose synthesis in hepatocytes and lipid degradation in adipocytes.[1,4–6,8] These roles of insulin-mediated IR activation in metabolism and homeostasis make them primary targets for metabolic disorders such as diabetes that affect millions of lives globally.[8]

Whereas, insulin has been well-studied for nearly a century, molecular level understanding of insulin binding to the IR and receptor activation has been a long standing goal in diabetes research.[8,11] Insulin engages the IR in a complex non 1:1 stoichiometry, and an elegant mathematical model of receptor activation based on the harmonic oscillator reproduced the associated kinetic properties.[12] A key structural breakthrough was achieved when McKern *et al.*, provided us with the first glimpse of the IR ectodomain and revealed an inverted ‘V’ like

conformation in the absence of insulin.[10] In 2013, the first high resolution structure of an insulin-IR complex revealed a single insulin molecule bound to the extracellular domain of the IR.[4] This structure showed that the high-affinity binding site of insulin lies between the leucine-rich L1 domain and the C-terminus of the α-domain (α-CT).[4] These structures were complemented by computational studies that emphasised the importance of these domains for insulin binding.[6,7,13–15] In 2018, single electron microscopy (EM) experiments have shown that activation of the insulin-IR complex involved transitioning from an inverted 'V' conformation to a T-like state that is thought to allow the transmembrane domains to come in proximity to each other and result in auto-phosphorylation and downstream signaling.[2] In the same year, Xu *et al.,* resolved the structure of the IGF-1R bound to its cognate ligand the insulin-like growth factor (IGF), which revealed both similarities and differences with the IR-insulin complex.[11] During the course of this work, other structures have demonstrated the presence of the hypothesised low-affinity insulin binding sites at the IR and shown insulin engagement at the IGF-1R, which have revealed the complexities of ligand-receptor engagement. [3,16,17]

This recent explosion in structural information for the IR and IGF-1R complexes led us to ask if approaches such as molecular dynamics (MD) and free energy perturbation (FEP) could elucidate the atom-level drivers of the observed functional profiles of known insulin analogs. These approaches, have started playing significant roles in unlocking molecular mechanisms that link structural changes with biological functions of proteins[18,19], and successful application to insulins could have major implications for structure-based design of novel analogs. In order to probe the utility and potential limitations of MD and FEP, we posed a series of questions – Firstly, can we shed light on the recognition of the endogenous ligands i.e. wild-type human insulin (WT-ins) and insulin-like growth factor 1 (IGF-1) at their respective cognate receptors? Additionally, using the same framework, can we then assess WT-ins engagement at the IGF-

1R to understand the molecular origins of the naturally occurring specificity? Secondly, can we study different well-characterized natural (disease-linked) and man-made insulin variants to understand potency and selectivity at a residue level? A century of research into insulins have led to the identification of a large number of variants with different metabolic and mitogenic potencies. An accurate recapitulation of the functional profiles across diverse insulin analogs would act as a demonstration of the generalizability of the method as a design tool. Finally, can the calculations be deconvoluted to separate out individual molecular-level contributions to the overall energetics of ligand binding? Such a deconstruction of ligand engagement with targets and off-targets would set the stage for the use of MD-FEP in rational structure-guided design of novel insulins and peptides in general.

Results

Analysis of the IR and IGF-1R structures used in this study:

IR and IGF-1R are functional homodimers that share a similar architecture in terms of their extra-cellular, transmembrane and kinase domains.[4,5] To date, EM and XRD have both provided high-resolution structures of both IR and IGF-1R in complex with their cognate peptide hormones, which broadly capture the homodimers in variations of the inverted ‘V’ or “T” conformations (Table S1). For this study, we chose the EM derived IR-insulin complex (PDB accession code: 6CE9) and the XRD structure of IGF-1-IGF-1R (PDB accession code: 5U8Q) as our reference structures.[2,11] It is to be noted that in this study, we have focused on the receptor HA sites.

In IR, the wild type insulin (WT-ins) symmetrically can bind to four binding sites (site 1, 1', 2 and 2') spanned over active dimeric insulin receptor.[3,17] Two of them (1 and 1') are designated as high affinity (HA) sites with the other two (2 and 2’) labelled as low affinity (LA) sites. The

labelling of the insulin binding sites in the IR as HA and LA sites stems from the large differences in insulin surface burial upon binding to these distinct sites (706 $Å^2$ in HA site and 394 $Å^2$ in LA site)[16], which in turn leads to differences in binding affinities. The L1 of one subunit and α-CT' and FnIII-1' of another subunit form the HA binding site.[2,4,16,20] From the experimental structures available for human insulin in complex with its cognate receptor, we observed that the insulin mainly adopts an "open" conformation in the HA sites when the overall receptor is in the "T" conformation (**Figure 1**). The two important structural features associated with the open conformation are the breaking of the two inter-chain disulfide bridges and the unwound conformation of C-termini of B-chain when compared to the closed conformation with which both inter-chain cysteine SS bonds are intact along with a secondary structure associated C-terminal end of the B chain.

The superposition of the structure of Insulin-IR and IGF-1-IGF-1R complexes revealed that the bound conformations of the cognate peptides and the ligand binding pockets share similarities but also show differences. Consistent with the observations correlating the overall receptor conformation with the state of the disulfide bonds in the bound peptides (**Table S1**), the IGF-1 is seen in a more "closed" state, whereas the insulin is seen in the "open" form.[3,16,17,20] Despite differences in the inter-chain disulfide bridges of the bound ligand, the overall conformations of insulin and IGF-1 in their respective binding sites share significant similarities with regards to overall superimposability of the A and B chains, and the similarity between the crucial GXXG motif and the C-terminus ends of the B chain (**Figure S1**).

MD simulations of the IR and IGF-1R ectodomains in complex with cognate and non-cognate ligands:

To assess peptide engagement with the HA sites of IR and IGF-1R, simulations of complexes with cognate and non-cognate ligands were performed using all atom MD of the fully hydrated systems at 300 K (Figure S2a). The starting coordinates for the simulations with IR were obtained from Scalpin *et al.*, [PDB accession code: 6CE9][2] and for IGF-1R from Xu *et al.*, [PDB accession code: 5U8Q].[11] Missing segments of both receptor structures were modelled using homology modelling. Models for non-cognate ligands such as insulin bound to IGF-1R were similarly obtained by homology modelling in order to maintain consistency of starting structures within the comparison framework. Unless otherwise stated, all ligands were modelled in the "open" state at the IR and in the "closed" state at the IGF-1R in line with the respective reference structures. After initial restrained equilibration, the entire system was simulated with complete flexibility.

The interaction framework provided by the initial structures was overall well-maintained throughout the trajectory and the system was largely stable [Figure 2: RMSD]. The low average fluctuation for each peptide ligand further elucidated the intrinsic stability of the bound conformation. In addition, the solvent accessible surface area (SASA) was also calculated for each ligand and the obtained averages were in line with those reported for the HA site of XRD/EM structures, further reiterating the stability of the complexes. However, differences in individual interaction patterns were also obtained for each simulation condition. As an indicator of ligand-receptor interaction strength, we initially calculated the enthalpic contribution for different complexes using the MM-GBSA method from the MD trajectories. (Table S2).

A key observation from the ΔH values was the significant drop in absolute values between the cognate and non-cognate complexes in the IGF-1R (Table S2). To investigate the structural origins of these observations we analysed intermolecular interactions with 4.5 Å distance cut-off around the ligand. In order to provide a simplistic framework for understanding the differences in intermolecular interactions we divided the WT-ins and IGF-1 (ligands) into sub-segments based on secondary structure elements (Figure 3). When an occupancy cut-off of 90% was applied to interactions within a 4.5 Å shell around the ligand, the analysis revealed that the cognate complexes maintain a similar interactions pattern between the ligand and the receptors when compared to the non-cognate complex (Figure 3). In the cognate complexes the highly stable interactions are dominated by those between the C-termini of both A- and B-chains (or equivalents in the case of IGF-1) of the respective ligands with the α-CT domain of the receptor. In terms of differences, whereas IGF-1 forms more extensive interactions with the α-CT domain of the IGF-1R with the equivalent of the insulin A-chain C-terminus, Insulin engages IR-α-CT more extensively with its B-chain. In the highly conserved aromatic B-chain or equivalent domain of both ligands, Insulin engages primarily with B24Phe, whereas IGF-1 engages with the equivalent B25-26 residues (Figure 3). Compared to both cognate complexes, binding of non-cognate insulin with IGF-1R resulted in distinct patterns of interaction. Whereas, interactions with the α-CT were still dominant, engagement with the Fn-III and L1 domains were also established that were different from the cognate IGF-1-IGF-1R complex (Figure 3).

Evaluation of insulin analogues at IR and IGF-1R:

To further understand receptor-ligand interactions, we aimed to quantify the effects of individual amino acid substitutions within the insulin molecule via evaluation of insulin analogs using FEP. Access to both IR and IGF-1R structures meant we could attempt to assess

quantification of both potency and selectivity of different insulin analogs. For consistency, FEP calculations were performed on the HA sites of the IR and IGF-1R using the same starting coordinates as our MD simulations [PDB accession codes: 6CE9 and 5U8Q]. These calculations were performed with spherical boundary conditions, which overall maintained the receptors in the ligand-bound states represented by the starting structures (**Figure 1, Figure S2a**).

All atoms within the sphere were fully flexible whereas residues outside the sphere were tightly restrained to their initial coordinates, thereby maintaining the overall conformation of the receptor (Figure S2a). Sphere radii were chosen for the IR and IGF-1R to ensure that all relevant receptor domains were included. No large fluctuations of structure were observed within the sphere during any of these simulations with RMSD values being < 0.3 Å. FEP schemes were devised for each analogue assessed in this study utilizing the fact that free energy is a state function and independent of path. The number of intermediate steps for each transformation was assigned in order to minimize hysteresis. In total 8.8 μs of simulations were performed, with an average of 140 ns used for each FEP calculation for each receptor. Relative binding free energies ($\Delta\Delta G_{bind}$) were calculated for each analog using a thermodynamic cycle with the free ligand in water acting as the reference state (Figure S2b). $\Delta\Delta G_{bind}$ with respect to WT-ins or equivalent was used to quantify the relative affinities of each molecule to the IR and IGF-1R in their respective conformations. In this work, insulins were divided into four classes, namely, commercially relevant analogs, disease mutants, truncated insulins, and a series of mutations at the position B10, which has attracted a lot of attention due to the potential for increased potency. In total, 20 MD/FEP calculations covering 13 mutations were performed. Overall, FEP calculations could correctly capture the direction of shift in 17 out of 20 or in 85% of cases (Figure 4, Table S3). In more than half of those cases, the deviation between

experiment and prediction was less than 1 kcal/mol (Figure 4, Table S3). As experimental functional profiles for analogs were in many cases obtained from multiple sources, the inner range (range of mean values) rather than the absolute spread of values including errors were considered to derive the range and mean of experimental $\Delta\Delta G_{bind}$, to assess the performance of FEP (Figure 4, Table S3).

Commercially interesting Insulin analogs:

Insulin analogs have been developed to specifically address different aspects of blood glucose control.[21,22] In this study, we have performed FEP calculations of Insulin Lispro, Insulin Aspart and an insulin carrying a mutation A21gly, which is present in Insulin Glargine. Lispro is a fast-acting insulin analog form Eli Lilly. The sequence of Lispro differs from WT-ins at the C-terminal end of the B-chain, involving a switch of residues B28Pro/B29Lys to B28Lys/B29Pro. Insulin Aspart carries a single amino-acid change from B28Pro to AspB28. Glargine is a long-acting insulin, carrying a glycine substitution at position A21, which we have characterized in this study, and the addition of two arginine molecules at the C-terminus of the B-chain, whose effects have not been evaluated here. The direction of shift in relative binding affinity ($\Delta\Delta G_{bind}$) values obtained from FEP calculations were in accordance with experiment in five of the six cases (83%) at IR and IGF-1R (Figure 4, Table S3).

For Aspart the $\Delta\Delta G_{bind}$ values at the IR was in line with its experimentally observed metabolic profile. At the IGF-1R, whereas the significant drop in mitogenic potency of Aspart was captured by the calculations, the magnitude was over-estimated. Further decomposition of $\Delta\Delta G_{bind}$ in IR reveal the loss of electrostatics which was compensated by the annihilation of the side chain carboxylic acid group. Structural analysis revealed a 90° rotation of the aromatic ring of B26 Tyr residue in the presence of Asp at B28 (Figure 5). In the case of IGF-1R, the

loss of electrostatics was not completely compensated by the annihilation of the aspartic acid residue, and thus resulted in an unfavourable $\Delta\Delta G_{bind}$ when compared to WT-ins at this receptor. Similarly, the calculated $\Delta\Delta G_{bind}$ for lispro at both IR and IGF-1R where in agreement with experimentally observed metabolic and mitogenic potencies, with a similar over-estimation observed at the off-target receptor. Importantly, the relatively larger decrease in mitogenic potency of Aspart compared to Lispro was captured by the calculation (Figure 4, Table S3). Finally, the $\Delta\Delta G_{bind}$ values for A21Gly indicated equipotency at both IR and IGF-1R for this mutant, which is largely in-line with experiments for its metabolic profile but could not capture the drop in mitogenic potency of this mutant (Figure 4, Table S3). Structural analysis also revealed largely compensatory interaction patterns observed with both Asparigine and Glycine as the C-terminal residue of the insulin A-chain.

Disease Mutants:

Missense mutations in insulin gene lead to three rare insulin mutants which cause hyperinsulinemia or hyperproinsulinemia.[23–25] and carriers were observed to develop insulin resistance that could lead to a diabetic state[23]. Three natural mutations that have been observed in different populations, namely insulin Wakyama (ValA3Leu), insulin Chicago (PheB25Leu), and insulin Los Angeles (PheB24Ser). These mutations confer reduced biological activity towards IR with Wakayama displaying the largest loss-of-function. FEP was used to probe the molecular drivers for the losses-of-function and the method accurately predicted the direction of shift for two of the three mutants (66%; Figure 4, Table S3).

The $\Delta\Delta G_{bind}$ value obtained for insulin Wakyama captured the loss-of-function observed at the IR. Decomposition of the $\Delta\Delta G_{bind}$ demonstrated that the annihilation of the larger alkyl chain of the Leucine residue restored favourable energies at the IR. Structural analysis further

revealed that a water-mediated interaction network connecting D496 (Fn-III domain), K703 & D707(α-CT) is lost when the naturally occurring valine is changed to the bulkier leucine (Figure 6a). Similarly, FEP captured the loss-of-function observed in the case of insulin Chicago. Structural analysis revealed subtle perturbations of the interaction network around B25Phe, when the aromatic phenylalanine residue is replaced with a Leucine (Figure 6b). However, the $\Delta\Delta G_{bind}$ for insulin Los Angeles indicated near equipotency for the variant, which was not in line with experiments (Table S3). However, it must be noted that the PheB24Ser variant still retains a significant portion of activity at the IR.[26] Structural analysis revealed that the bulky B24Phe is buried in a hydrophobic pocket, which when mutated to a smaller hydrophilic serine residue could lead to partial or incomplete solvation of the sub-pocket that may cause the incorrect prediction of $\Delta\Delta G_{bind}$ (Figure 6c).

B10 Substitutions:

In human insulin B10His plays a crucial role in the formation of insulin hexamers by the coordinating with zinc.[27,28] This position in the insulin sequence has also garnered significant interest as the substitution of the native histidine with an acidic residue (Asp or Glu) results in significant gain-of-function at both the IR and IGF-1R.[22,29] Analysis of B10 with static experimental structures have shed limited light on the molecular origin of these effects due to the absence of any direct interactions for B10 with the receptor in the IGF-1R. In the IR, HisB10 is seen proximal to an Arginine residue (510 from the FnIII-1' domain) but the ND1 atom is still 4.5 Å away from the basic nitrogen (NH2) of the Arg. FEP was used to probe the effects of a variety of experimentally-evaluated substitutions at the B10 position to better understand correlations between the nature of amino acid at this position on overall function. Overall, FEP could accurately predict the direction of shift in all eight cases (100%; Figure 4,

Table S3), with calculations on B10Val mutant highlighting the ability of the method to capture long-range effects on $\Delta\Delta G_{bind}$.

FEP was able to recapitulate the strong increase in HisB10Asp towards IR (Figure 4). Decomposition analysis revealed that a major portion of the gain in affinity towards the IR for B10Asp was driven by electrostatics which was only partially compensated by the annihilation of the imidazole ring atoms. Structural analysis of MD trajectories obtained from the simulation of B10Asp in complex with IR revealed that the Arg510 of FnIII-1' forms a salt bridge when this position is occupied by an aspartic acid (Figure 7c). At the IGF-1R, FEP captured the even larger increase in affinity that has been observed for the B10Asp variant, which was again driven by electrostatics. However, unlike in the IR structural analysis revealed no direct interactions with the receptor in either WT-ins or B10Asp-Ins. To understand the molecular origins of such a large gain in potency at the IGF-1R we turned to solvent interaction network analysis from large-scale MD simulations of the end states (B10His & B10Asp). This revealed a striking difference in the water mediated receptor ligand interactions between the WT-Ins and B10Asp at both IR and IGF-1R (Figure 7). The mutation of His to Asp at B10 resulted in large changes of first- and second-degree water mediated receptor-ligand contacts. These contacts are increased by ~300% at both receptors when B10 is occupied by aspartic acid compared to the native histidine (Figure 7, Table S4). Furthermore, these interactions span across different domains of the receptors, thereby bridging and stabilizing the entire ligand bound conformation in both IR and IGF-1R.

Additionally, Glendorf *et.al,* evaluated a series of ten different single amino acid substitutions at B10 position at IR, IGF-1R with respect to binding, metabolic and mitogenic potency. In this study, we have studied three additional representative amino acids mutations at B10,

namely B10A (small), B10V, (hydrophobic alkyl), and B10Q (larger polar). Whereas, their study considered DesB30-insulins, we performed the calculations with full-length insulin to maintain consistency, especially since a high-degree of correspondence between Des-B30 and full-length insulins were reported. The calculated $\Delta\Delta G_{bind}$ values with full-length insulins for these analogs largely correlated with the observed experimental affinities at both the receptors (Figure 4, Table S3). Additional structural analysis for these mutations are available in supplementary information (Figure S3). However, in the case of B10V, the calculated $\Delta\Delta G_{bind}$ suggested an increase in mitogenic potency that was not observed. Based on our water network analysis, which revealed the importance of long-range solvent networks on functional profile of B10 mutants, we hypothesized that there could be a subtle perturbation due to the deletion of B30. Impressively, when the same calculation was repeated with Des-B30-insulin, calculated $\Delta\Delta G_{bind}$ values revealed near equipotency which was in-line with experiment (Figure 4, Table S3). This result demonstrated the sensitivity of the method in capturing long-range effects.

Truncated Insulins:

Despite strong evolutionary conservation in the B-chain C-terminus for insulins, it was observed that the deletion of the residues (B26-B30) does not ablate its biological activity.[30–32] Conversely, it was seen that the use of specific residues at the new C-terminus (B25) along with amidation created large increases in potency compared to WT-ins. In this study, we evaluated three mutations at B25 for truncated [DesB26-30], WT-ins, based on the work of Casaretto *et al*.[32] The truncated mutants offered an opportunity to probe the generalizability of the FEP approach as deletion of the C-terminus represented a significant variation from WT-ins. In-line with the experimental work, the C-terminus of the B-chain was amidated in our

calculations. Overall, the FEP calculations could accurately capture direction of shift for two of three truncated systems (66 %; Figure 4, Table S3).

The $\Delta\Delta G_{bind}$ for the mutation B25His was strongly favorable in-line with experimental results (Figure4, Table S3). The decomposition of the $\Delta\Delta G_{bind}$ reveals that there is significant change in electrostatics compared to reference state, which was partially compensated by the annihilation of atoms. Structural analysis revealed the formation of a new network of water mediated, polar, and salt-bridge interactions between the receptor and ligand when histidine replaced phenylalanine as the C-terminus residue (Figure 8). Similarly, the increased efficacy of B25Tyr at the IR was also captured by our calculations (Table S3). Again electrostatics, was seen as the significant driver of more favorable $\Delta\Delta G_{bind}$ values. The $\Delta\Delta G_{bind}$ for B25Trp indicated an increase in affinity for this mutant which is not in line with the experiments. Decomposition of interactions have shown that a majority of the $\Delta\Delta G_{bind}$ originates upon the annihilation of the distal aromatic carbon atoms in the tryptophan, which is not compensated by electrostatics or further anhillation.

Discussion

A century on from the discovery of insulin, molecular-level understanding of its receptor engagement remains an important goal.[5,8] In this work, we have leveraged the recent increase in high-resolution structures[2,11] and used MD with FEP to elucidate the atomistic details of insulin recognition at the IR, its target and IGF-1R, its most important off-target. Three key results emerged from our evaluations. First, MD simulations provided us with a framework to capture, understand and map engagement of the endogenous ligands (WT-ins and IGF-1) with target (IR) and off-target (IGF-1R) receptors. Secondly, FEP was used to accurately recapitulate the functional effects for a variety of mutations spanning different segments of the

insulin sequence, demonstrating its potential for widespread application as a design tool. Finally, a deeper understanding of the links between molecular-level drivers and the observed functional profiles of different insulin variants were obtained, thus setting the stage for the use of MD-FEP in rational structure-guided design of novel insulins and peptides in general.

The conservation in receptor architecture between IR and IGF-1R (type II RTKs) leads to significant cross-affinity for the respective ligands.[33,34] Thus IGF-1R, a growth factor receptor whose out-of-turn activation may be oncogenic is a key off-target for any insulin analog.[35–37] Naturally, human insulin has an ~100-fold selectivity between IR and the IGF-1R and this was the starting point for trying to understand recognition of the ligand at both receptors.[38,39] Recent structural breakthroughs have resulted in multiple high-resolution structures of IR and IGF-1R and this served as a starting point for MD simulations of the respective cognate complexes and a non-cognate complex of WT-ins-IGF-1R.[2–4,11,16,17,20,33,34,40] The simulations revealed multiple commonalities in terms of ligand recognition at the two receptors, but also revealed key differences that may be exploited to rationally design selectivity. Importantly, MM-GBSA calculations captured the reduction in interactions (enthalpy) between cognate and non-cognate complexes. These are consistent with previous studies that have also shown that MM-GBSA can capture global differences in protein-protein interactions.[41] Analysis of close contacts, defined as a 4.5 Å sphere around the ligand, re-emphasized the importance of the α-CT in ligand recognition. The analysis of stable interactions, defined as having a lifetime of >90% revealed a network of key residue-ligand contacts, which were largely in-line with previous studies.[2,11] Interestingly, this analysis also revealed some key differences between the cognate and non-cognate complexes of WT-ins and IGF-1 with the receptors, setting the stage for a better understanding of ligand binding and specificity.

In order to further probe the contributions of individual amino acids within insulin towards potency via the IR, or mitogenic potency via IGF-1R, we quantified the free energy changes associated with natural and man-made mutations using FEP. In the last several decades, a number of insulin variants have been produced and characterized.[21,42–46] Two such approved insulin analogs Lispro and Aspart harbor changes at the C-terminal end of the insulin B-chain.[21,42,45] The FEP calculations accurately captured the near equipotency of Lispro and Aspart to WT-ins and significantly also captured the larger reduction in mitogenic potency that has been observed for both molecules. Structural analysis also revealed how mutations at one position in the sequence affects the conformations of neighbouring residues as seen in the case of Aspart. Whereas, the direction of shift was accurately captured at both target and anti-target, there was an overestimation seen for both molecules at the IGF-1R (Table S3). These mutations involved annihilation of charged amino acids such as lysines and aspartic acids and hence involved large changes in electrostatics, which has been shown previously to result in over-estimations due to incomplete reorganization of the micro-environment around the residues.[47] Structural and solvent-network analysis around these residues further confirmed significant reorganization of water networks occurred around the sites of annihilation. Another commercially interesting variant, the A21gly mutant was predicted by MD-FEP to be nearly equipotent at both IR and IGF-1R compared to WT-ins with minimal structural perturbation.[48] However, the mutation of asparagine to glycine has been experimentally shown to promote a significant reduction in mitogenic potency, which was not captured by our calculations.[45] However, our results are more in-line with other experiments, which have shown that alteration of Asn to Gly at A21 less significant effects on biological activity.[48,49]

Three major naturally occurring point mutations of human insulin have been identified and the presence of these increases the likelihood of developing diabetes.[23,25] The three variants,

named Wakayama (ValA3Leu), Los Angeles (PheB24Ser), and Chicago (PheB25Leu) result in loss-of-function at the IR.[23,25] An illustration of the power of structure-guided methods was seen in the calculations for Insulin Wakayama. Whereas, Valine to Leucine appears as a conservative mutation, there is a significant loss-of-function associated with this change to WT-ins. From the calculations, a disruption of direct and solvent-mediated interactions between the insulin variant and the L1 domain of the IR was revealed, which was also reflected in the unfavorable free energy change associated with this mutation. Similarly, a loss-of-function was predicted by our calculation for the Chicago variant, where a similar disruption of interaction networks between the receptor and ligand were identified. Interestingly, SerB24 that characterizes the Los Angeles variant has been shown to be one of the mutations that still retains a reasonable degree of metabolic potency. Our calculations indicated a small increase in potency at the IR, and did not capture the loss-of-function associated with this mutant. B24Phe resides in a buried hydrophobic pocket. Mutation to the smaller hydrophilic serine will probably result in a partially solvated pocket, where this residue is accommodated. The time-scales for our FEP calculations may not have been sufficient to allow for the required solvation of this buried pocket. Furthermore, since our calculations in this work was restricted to the HA pocket of the IR, it is possible that the more subtle loss-of-function associated with this change may originate from the LA pocket, or from a larger structural perturbation of the complex. Overall, the simulations managed to shed significant light on the molecular drivers of loss-of-function associated with disease-associated mutants of insulin.

The B10 position within the insulin sequence has attracted significant attention due to the super-potency (~375% of human insulin) associated with the mutation of the natural histidine at this position to acidic residues such as aspartic or glutamic acid.[27,28] However, the increase in IGF-1R potency is even stronger than that seen at the IR, limiting the use of these analogs

due to safety concerns.[45,48,50] Hence, understanding the structural drivers of the potency increase could prove important in leveraging the potential of B10 mutants. The MD-FEP calculations managed to capture the significant increase in potency associated with the B10Asp variant at the IR and importantly recapitulated the even stronger effect at the IGF-1R. Structural analysis at the IR, revealed the importance of interaction with R510 of FnIII-1' domain on the increased potency of B10Asp. Surprisingly, B10 was found to be completely solvent exposed at the IGF-1R with no direct interaction with the receptor. However, analysis of the solvent networks around the residue, revealed that the change from histidine to aspartic acid promoted a strong network of water-mediated interactions including inter-domain solvent bridges that stabilized the overall bound conformation at the IR and even more so at the IGF-1R [Figure 7]. The solvent network analysis demonstrates the power of MD simulations, where the use of explicit water molecules can capture important effects not readily apparent from static structures or even continuum solvation methods. The strength of FEP was further emphasized in the case of the B10V mutation. Whereas, calculations for this mutant with full-length insulin predicted an increased mitogenic potency, the same was not observed in experiments with Des-B30-B10Val, despite DesB30-insulins having highly similar functional profiles to WT-ins.[27] The calculations when repeated with Des-B30 insulin immediately recapitulated the experimental results revealing the ability of the method to capture subtle, long-range effects of specific amino-acid mutations.

Fascinatingly, previous studies have shown that deleting the last five highly conserved residues at the C-terminus of the insulin B-chain still results in a molecule with significant potency.[31,32] In fact, changes in the new C-terminus (B25) of these truncated molecules, provide variants with increased potency at the IR compared to WT-ins.[31,32,51] Despite, the removal of the C-terminus of the B-chain, a change of five amino acids, FEP was able to accurately reproduce

the experimental functional profiles in a majority of cases, showing its versatility and generalizability. The increased potency associated with Histidine and tyrosine as C-terminus was captured by our calculations along with an understanding of its molecular origins, namely an altered polar interaction network around the new residues. Overall, the ability of the method to work across different types of insulin variants including molecules carrying significant changes in sequence length points to the power of MD-FEP as a generalized tool for the design of novel insulins and beyond.

Conclusion

In this study, we have demonstrated the utility of MD and FEP in elucidating the molecular drivers of insulin recognition at its target and off-target receptors . The ability of these methods to probe, recapitulate and understand the effects of a variety mutations on functional profiles on insulin analogs lays the foundation for its use in rational structure-based design of novel variants. The accuracy observed across different analog types and amino-acid mutations demonstrated the generalizability of our approach, which acts as the centerpiece of our Sq-PROT platform, for rational peptide/biologics design.

Methods

Molecular Dynamic Simulations:

The MD simulations were performed with the experimentally obtained structures for IR/IGF-1R (PDB accession code: 6CE9; 5U8Q).[2,11] Initial coordinates for the receptor-ligand complexes were obtained directly from the experimental structures for the cognate systems with missing sections modelled based on available templates, or from homology models for non-cognate conditions. In the first step, fully hydrated complexes were equilibrated in AMBER[52] using the AMBER14ffSB[53] all atom force field, and TIP3P[54] waters. The

protonation state for ionizable residues were set to their most probable state and histidine tautomers were assigned by manual inspection. During equilibration, restraints were placed on protein atoms, which were removed in a series of equilibration steps spanning a total of 10 ns. After equilibrating the system, a production run of 50 ns was performed at 300 K temperature with a non-bonded cut-off of 10 Å . The SHAKE algorithm was used to constrain bond lengths during the simulation.[55] Long range electrostatics were treated with Ewald summation method. [56] The Langevin thermostat was used to maintain the temperature on the system, while Berendsen barostat maintained pressure at 1 Bar with a coupling of 2 ps.[57,58]

Free Energy Perturbation (FEP) Method:

The free energy calculations were performed using the equilibrated systems with spherical boundary conditions in the program Q.[59] The simulations were performed at constant temperature of 310 K in a spherical droplet of 27 Å centred on the Cδ atom of B15Leu at IR, whereas the sphere radius was set to 30 Å at the IGF-1R to include key residues and domains. Atoms within the sphere were fully flexible, while those lying outside the sphere were tightly restrained to their initial coordinates. Ionizable residues within the sphere were protonated according to to their most probable states, and those close to the sphere edge were neutralized. The protonation state for histidines within the sphere were set by manual inspection. A non-bonded cut-off of 10 Å was used for all atoms except the amino acid residue being mutated for which there is no cutoff. The SHAKE algorithm was used to constraint all the bonds and angles of protein atoms and the sphere surface were subjected to radial and polarization restraints according to the SCAAS model.[55,60,61] Long range electrostatics were treated with local reaction field (LRF) method.[62] The time step was set to 1 fs and non-bonded pair lists were updated every 25 steps.

The relative binding free energy for a pair of residues was calculated in series of steps using MD/FEP, which in general are, i) The transformation of partial charges and (ii) combined transformation of Lennard-Jones (LJ) and parameters involving covalent bonds in FEP calculations. If multiple atoms were annihilated, separate MD/FEP calculations were carried out to remove these in step-wise manner. A softcore potential was introduced for the atom in a first step and followed by removal of the resulting van der Waal's potential. The total free energy is calculated as sum of the results obtained in each step. Each MD/FEP calculation was divided into n intermediate steps that were equilibrated separately. The potential (U_m) defining each state was a linear combination of energy functions describing the start-A and end point (B) of the transformation.

$$U_m = (1-\lambda_m)U_A + \lambda_m U_B$$

where λ_m varies from 0 to 1. The transformations involving partial charges were performed using up to 36 λ steps and the number of λ values used to transform LJ and bonded parameters ranged from 42 to 63 steps. The ligand-receptor complex was typically equilibrated for 1.76 ns at each λ value. The harmonic restraints were released on ligand-receptor complexes in several steps and the temperature was increased in a series of steps to 310 K. After equilibration, the unrestrained simulation for the complex was typically carried out for 0.5 ns for each λ value. The ligand in water consisted of the entire insulin molecule and was typically equilibrated for 0.76 ns followed by 0.25 ns of unrestrained simulation for each λ. The same sphere radii were maintained between the respective receptor and water simulations. The difference in free energy between initial state (A) and final state (B) was calculated by summing up the free energy differences of the n intermediate states using the relation

$$\Delta G_{A\rightarrow B}^{FEP} = -kT \sum_{m=1}^{n-1} \ln \langle e^{\frac{-(U_{m+1}-U_m)}{kT}} \rangle_m$$

Where $\langle ... \rangle_m$ represents an ensemble average on the potential U_m which is calculated from the MD simulations.[63] The uncertainty of a transformation was quantified as the difference in free energy obtained by applying FEP formula in the forward and reverse direction and was minimized by increasing the number of l values or simulation length.

MM-GBSA:

The MM-GBSA method is used to determine the binding free energies of protein-ligand complexes.[64] This method uses molecular mechanics, Born model to compute free energies as sum of individual energy components.[65] For a ligand (L) binding to a receptor (R-) -

$$L + R \leftrightarrows LR$$

The free energy of the above complex can be calculated using MM-GBSA using below steps,

$$G = E_{bond}+E_{el}+E_{vdw}+G_{pol}+G_{np}-TS$$

Where E_{bond} are energy from bonded terms, E_{el} are electrostatic energy, & E_{vdw} is the van der Waals energy. The above terms are obtained from molecular mechanics calculations (MM). G_{pol} and G_{np} are polar and non-polar terms of the solvation free energy. G_{pol} is obtained by using generalized born model, whereas G_{np} is obtained from the solvent accessible surface area (SASA). The last term of the equation is absolute temperature (T) and entropy which is estimated from normal-mode analysis. The binding free energy is calculated using the below equation.

$$\Delta G_{bind} = \Delta E_{MM} + \Delta G_{solv} + \Delta G_{SA}$$

where ΔE_{MM} is the difference in energy between the receptor and sum of the energies of ligand and unliganded receptor. ΔG_{solv} is the difference is GBSA solvation energy of the receptor and sum of the solvation energies of ligand and receptor. ΔG_{SA} is the difference in surface energy of the receptor complex and sum of the surface energies of ligand and receptor alone. The MM-GBSA calculations were performed on MD trajectories obtained using Amber as described above for cognate and non-cognate complexes.

Homology modelling:

Homology models for the non-cogate complex of WT-ins in IGF-1R were generated using MODELLER v9.25.[66] In total, ten models were generated for the non-cognate complex using the IGF-1-IGF-1R complex as template (PDB accession code: 5U8Q).[11] The overall, conformation of the modelled WT-ins was maintained in the same state as that of IGF-1 in the IGF-1R. The selection of a model to be used for further calculations were based on DOPE score[67] and visual inspection. In a similar manner, missing segments of the receptors, such as a segment of a-CT in IGF-1R was generated using MODELLER v9.25 using suitable templates from available structures.

Water(solvent)-interaction network analysis:

To assess the solvent and solvent-mediated interaction networks around a selected residue a distance-based analysis scheme was devised. Initially, snapshots of the MD trajectories were saved as PDB structures. Hydrogens atoms were removed from the snapshots obtained from the MD trajectories and the residue of interest was provided as an input. Only side chain heteroatoms of the defined residue on interest were considered for the analysis. Polar contacts with either water or protein were defined within a 3.5 Å radius. First order contacts are those polar contacts with the receptor that were within 3.5 Å of the residue of interest (P0), and waters within the same cutoff (W0). Using the W0 waters as centers, the analysis was repeated. A receptor polar contact within 3.5 Å of any W0 water was defined as a second order contact (P1) and any water molecule within the same radius as a second-order water contact (W1), The process was repeated with W1 waters as centers resulting in a list of third-order receptor (P2) or water (W2) contacts. Average values of W0, P0, W1, P1, W2 and P2 for any residue of interest were obtained by averaging across the number of considered snapshots.

ACKNOWLEDGMENTS

We thank Sudha Ravishankar and Sampath Ranganathan for their valuable inputs in the preparation of this manuscript.

Author Contributions

M.M.S., R.C., and A.R. designed the research project. M.M.S., R.C., and A.R. performed and analysed the MD simulations and FEP calculations. All authors contributed to the subsequent discussion of the results. All authors contributed to the writing of the manuscript.

Competing Interests

M.M.S., R.C., A.K., M.C., and A.R. are/were employed by Sekkei Bio Private Limited and/or hold shares in Sekkei Bio Private Limited. This does not alter the authors adherence to the journals policies on sharing data and materials. There are no patents belonging to Sekkei Bio Private Limited relating to material pertinent to this article. All molecules considered in this manuscript are known molecules present in the public domain.

REFERENCES

1. Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).

2. Scapin, Giovanna, et al. Structure of the insulin receptor–insulin complex by single-particle cryo-EM analysis. *Nature* **556,** 122-125 (2018).

3. Gutmann, T. *et al.* Cryo-EM structure of the complete and ligand-saturated insulin receptor ectodomain. *J. Cell. Biol.* **219**, 210-236 (2019).

4. Menting, John G., et al. How insulin engages its primary binding site on the insulin receptor. *Nature* **493**, 241-245 (2013).

5. De Meyts, Pierre. Insulin and its receptor: structure, function and evolution. *Bioessays* **26,** 1351-1362 (2004).

6. Papaioannou, A., Kuyucak, S. & Kuncic, Z. Computational study of the activity, dynamics, energetics and conformations of insulin analogues using molecular dynamics simulations: application to hyperinsulinemia and the critical residue B26. *Biochem. Biophys. Rep.***11**, 182–190 (2017).

7. Žáková, L. *et al.* Structural integrity of the B24 site in human insulin is important for hormone functionality. *J. Biol. Chem.* **288**, 10230–10240 (2013).

8. Mathieu, C., Martens, P. J., & Vangoitsenhoven, R.. One hundred years of insulin therapy. *Nat. Rev. Endocrinol* ***17***, 715-725 (2021).

9. De Meyts, P. & Whittaker, J. Structural biology of insulin and IGF1 receptors: implications for drug design. *Nat. Rev. Drug. Discov.* **1**, 769–783 (2002).

10. McKern, N. M. *et al.* Structure of the insulin receptor ectodomain reveals a folded-over conformation. *Nature* **443**, 218–221 (2006).

11. Xu, Y. *et al.* How ligand binds to the type 1 insulin-like growth factor receptor. *Nat Commun* **9**, 821 (2018).

12. Kiselyov, V. V., Versteyhe, S., Gauguin, L. & De Meyts, P. Harmonic oscillator model of the insulin and IGF1 receptors' allosteric binding and activation. *Molec. Sys. Biol.* **5,** 243 (2009).

13. Papaioannou, A., Kuyucak, S. & Kuncic, Z. Molecular dynamics simulations of insulin: elucidating the conformational changes that enable its binding. *PLOS ONE* **10**, e0144058 (2015).

14. Zoete, V., Meuwly, M. & Karplus, M. A Comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. *J. Mol. Biol.* **342**, 913–929 (2004).

15. Islam, M. A., Bhayye, S., Adeniyi, A. A., Soliman, M. E. S. & Pillay, T. S. Diabetes mellitus caused by mutations in human insulin: analysis of impaired receptor binding of insulins Wakayama, Los Angeles and Chicago using pharmacoinformatics. *J. Biomol. Struct. Dyn.* **35**, 724–737 (2017).

16. Uchikawa, E., Choi, E., Shang, G., Yu, H. & Bai, X. Activation mechanism of the insulin receptor revealed by cryo-EM structure of the fully liganded receptor–ligand complex. *eLife* **8**, e48630 (2019).

17. Nielsen, J. *et al.* Structural investigations of full-length insulin receptor dynamics and signalling. *J. Mol. Biol.* **434**, 167458 (2022).

18. Ranganathan, A., Dror, R. O. & Carlsson, J. Insights into the role of Asp79$^{2.50}$ in β2 adrenergic receptor activation from molecular dynamics simulations. *Biochemistry* **53**, 7283–7296 (2014).

19. Mey, Antonia SJS, et al. Best practices for alchemical free energy calculations. *Living journal of computational molecular science* **2**, (2020).

20. Menting, J. G. *et al.* Protective hinge in insulin opens to enable its receptor engagement. *Proc. Natl. Acad. Sci. U S A* **111**, E3395–E3404 (2014).

21. Hirsch, I. B. Insulin Analogues. *N. Engl. J. Med.* **352**, 174–183 (2005).

22. Drejer, K. The bioactivity of insulin analogues from in vitro receptor binding to in vivo glucose uptake. *Diabetes/Metab. Res. Rev.* **8**, 259–285 (1992).

23. Nishi, M. & Nanjo, K. Insulin gene mutations and diabetes. *J. Diabetes Investig*. **2**, 92–100 (2011).

24. Tager, H., et al. A structurally abnormal insulin causing human diabetes. *Nature* **281,** 122-125 (1979).

25. Shoelson, S. *et al.* Three mutant insulins in man. *Nature* **302**, 540–543 (1983).

26. Shoelson, S. E., Lu, Z. X., Parlautan, L., Lynch, C. S. & Weiss, M. A. Mutations at the dimer, hexamer, and receptor-binding surfaces of insulin independently affect insulin-insulin and insulin-receptor interactions. *Biochemistry* **31,** 1757-1767 (1992).

27. Glendorf, T. *et al.* Systematic evaluation of the metabolic to mitogenic potency ratio for B10-substituted insulin analogues. *PLOS ONE* **7**, e29198 (2012).

28. Brange, J. *et al.* Monomeric insulins obtained by protein engineering and their medical implications. *Nature* **333**, 679–682 (1988).

29. Schwartz, G. P., Burke, G. T., & Katsoyannis, P. G.. A superactive insulin:[B10-aspartic acid] insulin (human). *Procs. Nat. Aca. Sci. U. S. A.* ***84,*** 6408-6411 (1987).

30. Nakagawa, S. H. & Tager, H. S. Role of the phenylalanine B25 side chain in directing insulin interaction with its receptor. Steric and conformational effects. *J. Biol. Chem.* **261**, 7332–7341 (1986).

31. Fischer, W. H., Saunders, D., Brandenburg, D., Wollmer, A. & Zahn, H. A shortened insulin with full in vitro potency. *Chem. Hoppe-Seyler* **366**, 521–526 (1985).

32. Casaretto, M. *et al.* Shortened insulin with enhanced in vitro potency. *Biol. Chem. Hoppe-Seyler* **368**, 709–716 (1987).

33. Menting, J. G. *et al.* Structural congruency of ligand binding to the insulin and insulin/type 1 insulin-like growth factor hybrid receptors. *Structure* **23**, 1271–1282 (2015).

34. Croll, T. I. *et al.* Higher-resolution structure of the human insulin receptor ectodomain: multi-modal inclusion of the insert domain. *Structure* **24**, 469–476 (2016).

35. Pollak, M. Insulin and insulin-like growth factor signalling in neoplasia. *Nat. Rev. Cancer*. **8**, 915–928 (2008).

36. Buck, E. *et al.* Compensatory insulin receptor (IR) activation on inhibition of insulin-like growth factor-1 receptor (IGF-1R): rationale for cotargeting IGF-1R and IR in cancer. *Mol. Cancer Ther.* **9**, 2652–2664 (2010).

37. Renehan, A. G., Frystyk, J. & Flyvbjerg, A. Obesity and cancer risk: the role of the insulin–IGF axis. *Trends Endocrinol. Metab.* **17**, 328–336 (2006).

38. Křížková, K. *et al.* Insulin–insulin-like growth factors hybrids as molecular probes of hormone:receptor binding specificity. *Biochemistry* **55**, 2903–2913 (2016).

39. Hexnerová, R. *et al.* Probing receptor specificity by sampling the conformational space of the insulin-like growth factor II c-domain . *J. Biol. Chem.* **291**, 21234–21245 (2016).

40. Li, J., Choi, E., Yu, H. & Bai, X. Structural basis of the activation of type 1 insulin-like growth factor receptor. *Nat. Commun.* **10**, 4567 (2019).

41. Sindhikara, D. *et al.* Automated design of macrocycles for therapeutic applications: from small molecules to peptides and proteins. *J. Med. Chem.* **63**, 12100–12115 (2020).

42. Gammeltoft, S. *et al.* Insulin aspart: a novel rapid-acting human insulin analogue. *Expert. Opin. Investig. Drugs.* **8**, 1431–1442 (1999).

43. Heinemann, L., Starke, A. A. R., Heding, L., Jensen, I. & Berger, M. Action profiles of fast onset insulin analogues. *Diabetologia* **33**, 384–386 (1990).

44. Bolli, G. B., et al. Insulin analogues and their potential in the management of diabetes mellitus. *Diabetologia* **42**, 1151-1161 (1999).

45. Kurtzhals, P., Schäffer, L., Sørensen, A., Kristensen, C., Jonassen, I., Schmid, C., & Trüb, T. (2000). Correlations of receptor binding and metabolic and mitogenic potencies of insulin analogs designed for clinical use. *Diabetes*, *49*, 999-1005 (2000)..

46. Bolli, G. B. & Owens, D. R. Insulin glargine. *The Lancet* **356**, 443–445 (2000).

47. Simonson, T., Carlsson, J. & Case, D. A. Proton binding to proteins: pKa Calculations with Explicit and Implicit Solvent Models. *J. Am. Chem. Soc.* **126**, 4167–4180 (2004).

48. Sommerfeld, M. R. *et al.* In vitro metabolic and mitogenic signaling of insulin glargine and its metabolites. *PLOS ONE* **5**, e9540 (2010).

49. Fawcett, J., Tsui, B. T., Kruer, M. C. & Duckworth, W. C. Reduced action of insulin glargine on protein and lipid metabolism: possible relationship to cellular hormone metabolism. *Metabolism* **53**, 1037–1044 (2004).

50. Falck HANSEN, B. *et al.* Sustained signalling from the insulin receptor after stimulation with insulin analogues exhibiting increased mitogenic potency. *Biochemical Journal* **315**, 271–279 (1996).

51. Kurapkat, G. *et al.* The solution structure of a superpotent B-chain-shortened single-replacement insulin analogue. *Protein Sci.* **8**, 499–508 (1999).

52. Case, David A., et al. The Amber biomolecular simulation programs. *J. Comput. Chem.* **26,** 1668-1688 (2005).

53. Maier, J. A. *et al.* ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory. Comput.* **11**, 3696–3713 (2015).

54. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

55. Kräutler, V., van Gunsteren, W. F. & Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *J. Comput. Chem.* **22**, 501–508 (2001).

56. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

57. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

58. Ryckaert, J.-P., Ciccotti, G. & Berendsen, H. J. C. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **23**, 327–341 (1977).

59. Bauer, P. *et al.* Q6: A comprehensive toolkit for empirical valence bond and related free energy calculations. *SoftwareX* **7**, 388–395 (2018).

60. Marelius, J., Kolmodin, K., Feierberg, I. & Åqvist, J. Q: a molecular dynamics program for free energy calculations and empirical valence bond simulations in biomolecular systems. *J. Mol. Graph. Model.* **16**, 213–225 (1998).

61. King, G. & Warshel, A. A surface constrained all-atom solvent model for effective simulations of polar solutions. *J. Chem. Phys.* **91**, 3647–3661 (1989).

62. Lee, F. S. & Warshel, A. A local reaction field method for fast evaluation of long-range electrostatic interactions in molecular simulations. *J. Chem. Phys.* **97**, 3100–3107 (1992).

63. Kollman, Peter. Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* **93,** 2395-2417 (1993).

64. Kollman, P. A. *et al.* Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum Models. *Acc. Chem. Res.* **33**, 889–897 (2000).

65. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert. Opin. Drug. Discovery.* **10**, 449–461 (2015).

66. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234**, 779–815 (1993).

67. Shen, M. Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, *15*(11), 2507-2524.

68. Gutmann, T., Kim, K. H., Grzybek, M., Walz, T. & Coskun, Ü. Visualization of ligand-induced transmembrane signaling in the full-length human insulin receptor. *J. Cell. Biol.* **217**, 1643–1649 (2018).

69. Torlone, E., Di Cianni, G., Mannino, D. & Lapolla, A. Insulin analogs and pregnancy: an update. *Acta. Diabetol.* **46**, 163–172 (2009).

70. Pollak, M. & Russell-Jones, D. Insulin analogues and cancer risk: cause for concern or cause célèbre? *Int. J. Clin. Pract.* **64**, 628–636 (2010).

Figures Main Text:

MD simulations of the IR and IGF-1R ectodomains in complex with cognate and non-cognate ligands:

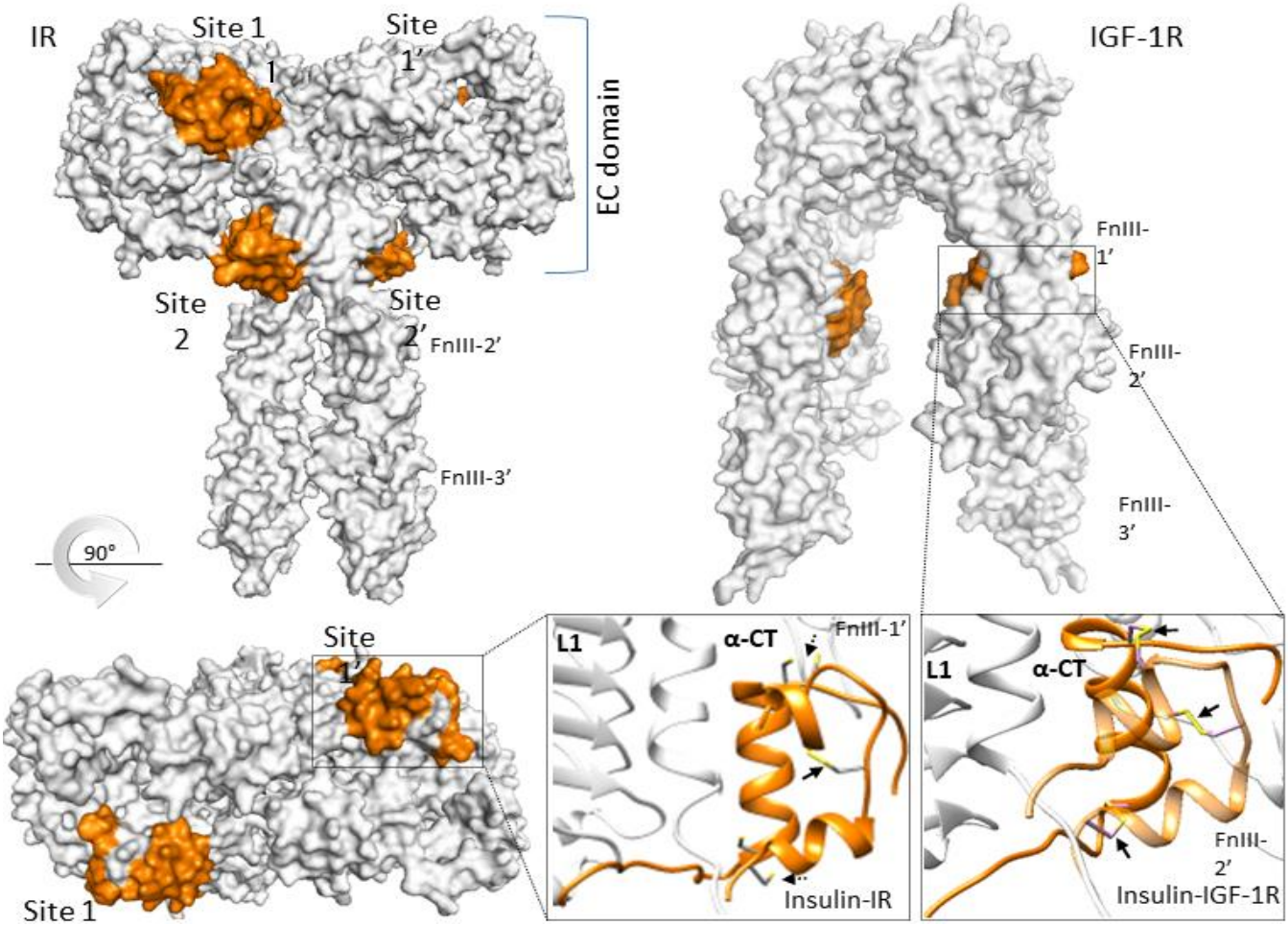

Figure 1. **Structural analysis of experimentally obtained structures of IR and IGF-1R:** Top panel - The experimentally obtained structures of the IR & IGF-1R receptors used in this work are shown in the surface representations with their respective 'T' and inverted 'V' shaped conformations. The binding sites for both the receptors are highlighted in orange. For IR, the high and low affinity sites are labeled as 1/1' and 2/2', respectively. Bottom panel – A zoom in of the high-affinity sites of both receptors are shown with the insulin molecule bound (represented as an orange cartoon). Arrows depict the intact inter- and intra-chain disulfide bonds in the non-cognate complex and the broken inter-chain bridges in the cognate complex.

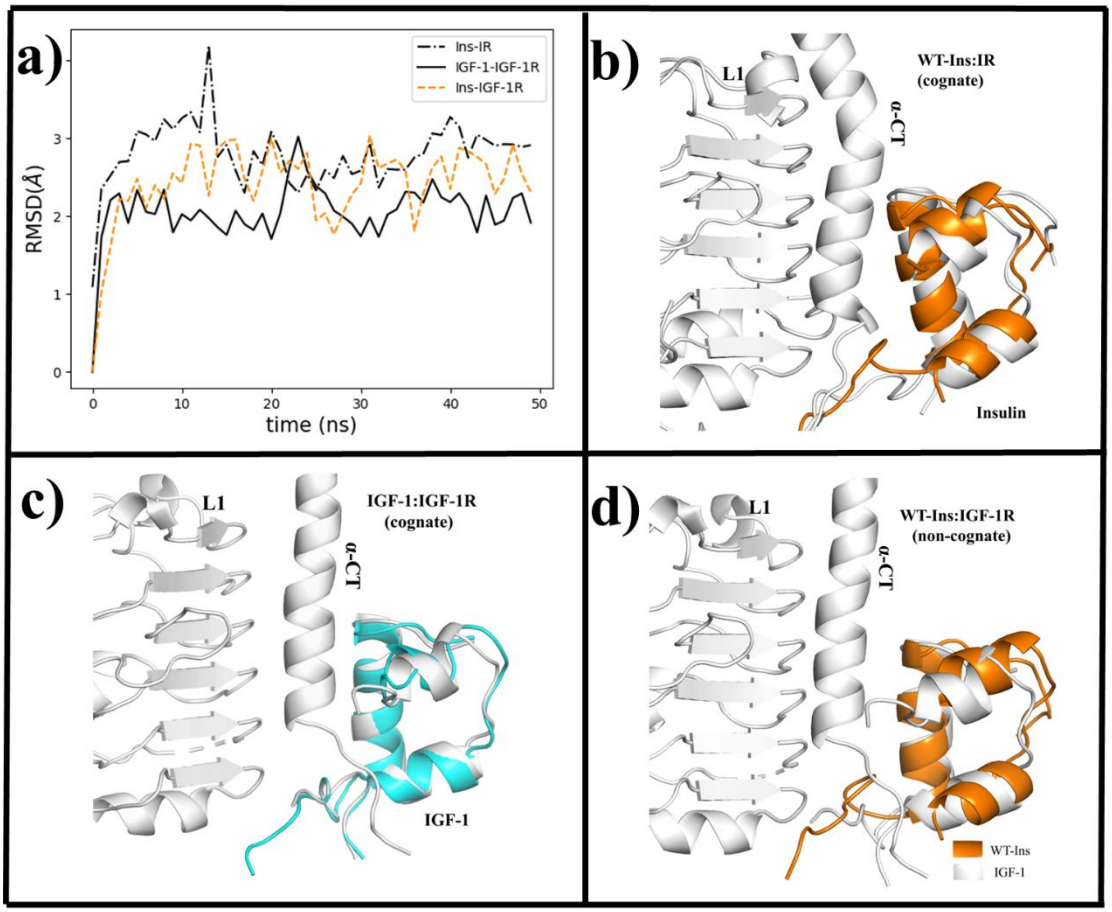

Figure 2: Assessment of stability of insulin-cognate and non-cognate complexes from MD simulations with the starting coordinates as reference a) Ca-RMSD of the three simulation conditions with respect to the initial structure, insulin-IR (black dashdot), IGF-1-IGF-1R (black, solid line), and insulin-IGF-1R (orange, dashed line), b) insulin-cognate complex: the insulin starting structure (grey cartoon) is shown in alignment with the average structure of insulin (orange cartoon) and receptor (grey cartoon) obtained from MD simulations c) IGF-1-cognate complex- the IGF-1 starting structure (grey cartoon) is shown in alignment with the average structure of IGF-1 (cyan cartoon) and IGF-1R (grey cartoon) obtained from MD simulations d) insulin-non cognate complex: the IGF-1 starting structure (grey cartoon) is shown in alignment with the average structure of insulin (orange cartoon) and IGF-1R (grey cartoon) obtained from MD simulations

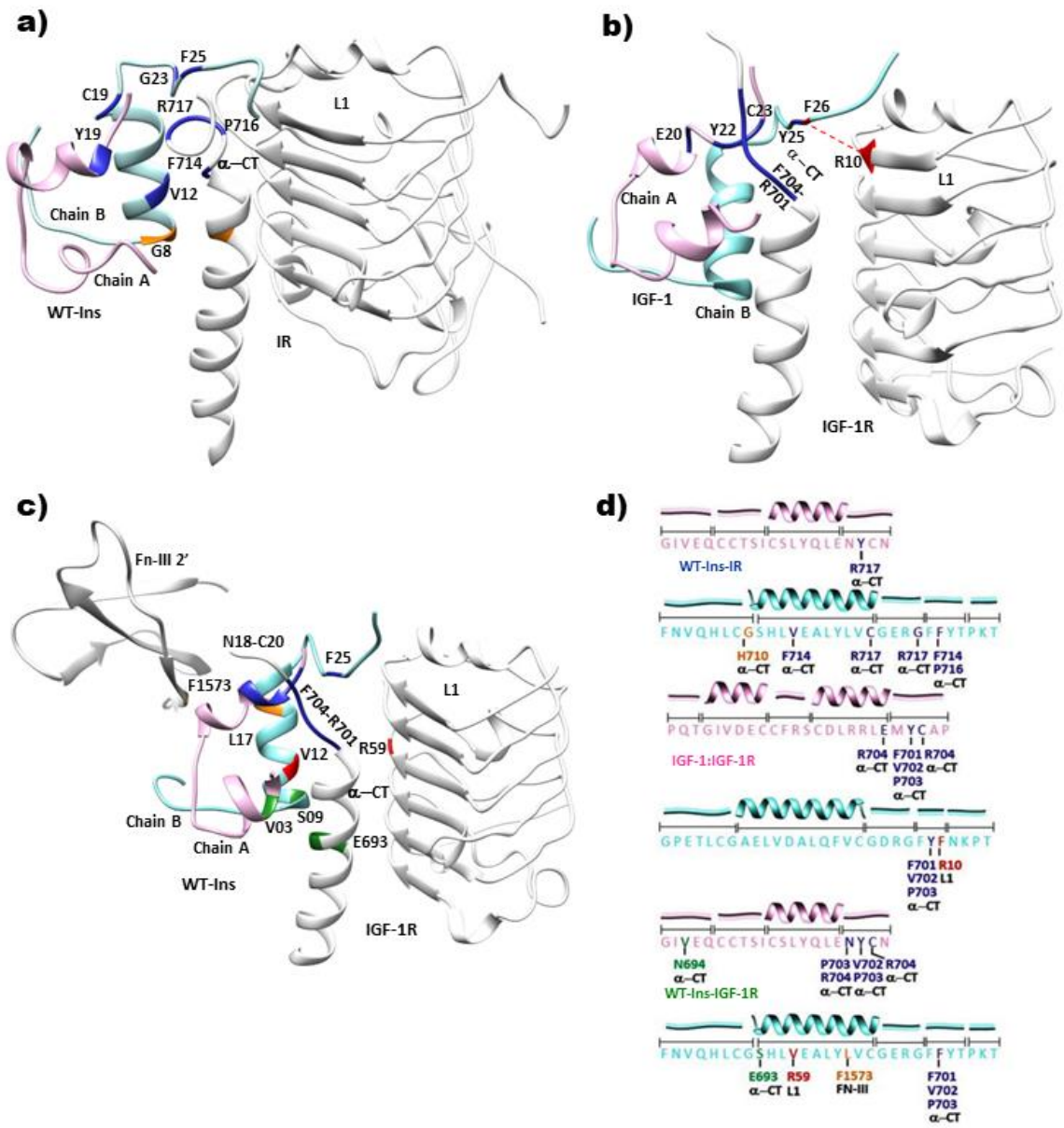

Figure 3: Analysis of stable contacts of WT-Ins in IR and IGF-1R receptors with more than 90% lifetime obtained from MD simulations. Ligand-receptor contacts were defined using a distance cutoff of 4.5 Å from any ligand residue. **a-c**) Represents the key contacts in the cognate and non-cognate complexes. The respective receptors in each complex IR or IGF-1R are represented as grey cartoons, the chain-A or chain-A equivalent of the respective ligands WT-ins or IGF-1: cyan and the chain-B or chain-B equivalent as Mauve. The interacting residues from ligand and receptor are identically color-coded domain-wise and domain names/residue names are further highlighted in text **a**) shows the interactions of WT-Ins with IR (cognate).

b) shows the interactions in IGF-1-IGF-1R (cognate). An additional dashed line (red color) shows the interaction of R10 from L1 domain with F26 of chain-B equivalent of the ligand; **c**) WT-Ins- IGF-1R (non-cognate); **d**) shows the sequence view of WT-Ins and the resolved portion of IGF-1. The interactions of ligand residues with specific receptor residues in the cognate and non-cognate complexes, namely, WT-Ins-IR (top), IGF-1:IGF-1R (middle), WT-Ins-IGF-1R (bottom) are shown in text. For convenience IGF-1 is shown as two chains to allow for easy comparison. The interacting residues are color-coded domain-wise using the same scheme as used in the structures a-c.

Evaluation of insulin analogs at IR and IGF-1R:

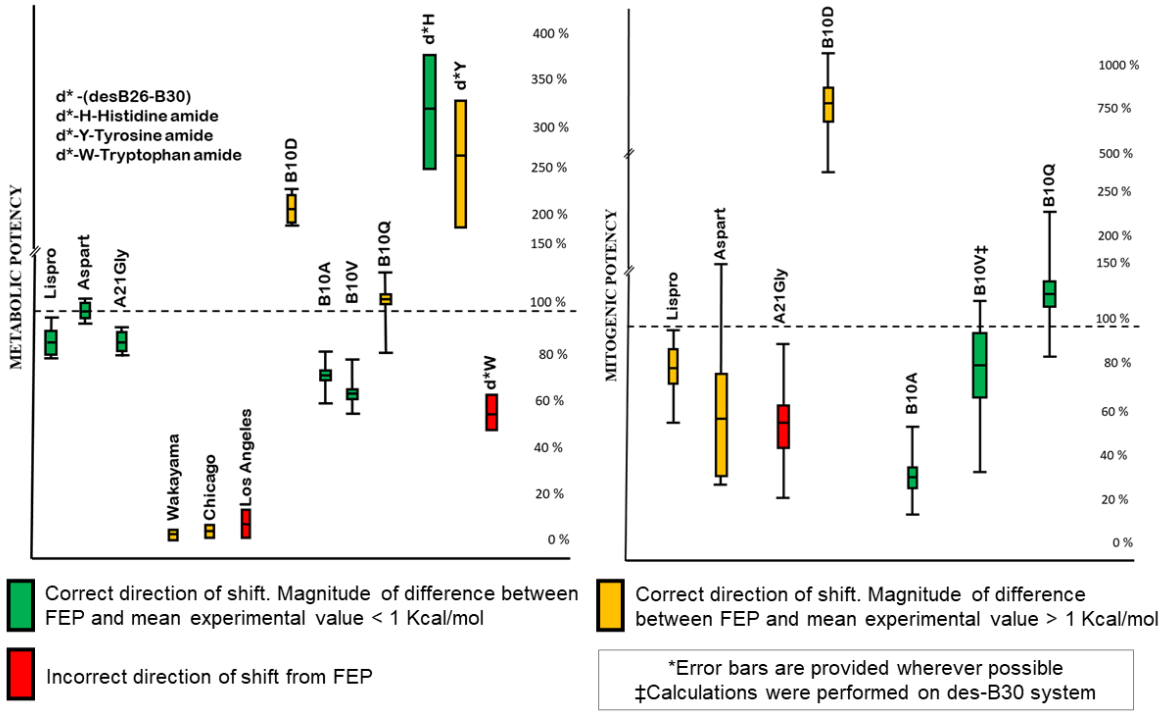

Figure 4: Summarizing the correspondence obtained between calculated values from FEP calculations & the relevant experimentally obtained values for all the insulin variants evaluated in this study. The relative binding affinities $\Delta\Delta G_{bind}$ obtained from the FEP calculations at IR were compared with experimental metabolic potencies (left) and $\Delta\Delta G_{bind}$ obtained from the FEP calculations at IGF-1R from experimental mitogenic potencies. The metabolic and mitogenic potencies of mutations are reported to relative to WT-ins whose potency values are set to 100%. As multiple experimental values were available for many of the analogs, only the spread of mean values were considered to evaluate the accuracy of an FEP calculation. The spread of the mean values are captured by the width of the individual bar. The absolute spread including the experimentally reported errors are shown using the error-bar representations. A horizontal line in the bar represents the mean value (as mean of means). The numerical values and full list of references are provided in Table S3. For the B10 mutant series, the value shown for B10V at IGF-1R was re-calculated using Des-B30-WT-ins to evaluate the effect of this

change. All other B10 calculations were performed using full-length WT-ins. This is discussed extensively in text.

Commercially interesting Insulin analogs:

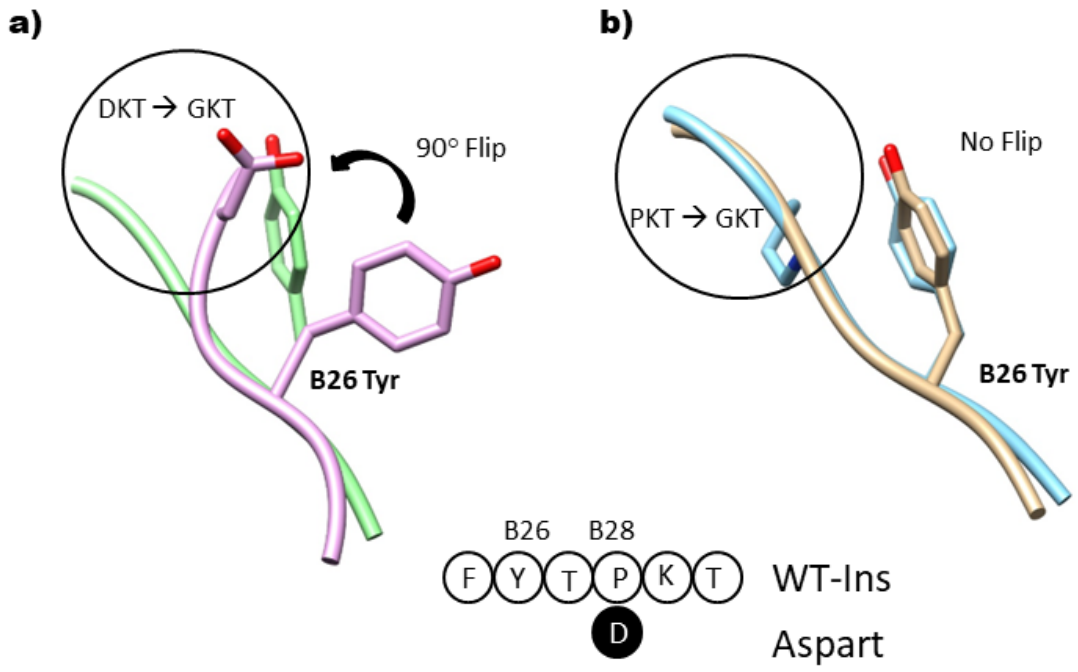

Figure 5: Analysis of Aspart showing the structural changes accompanying the transformation from WT-insulin to the mutant aspart a) showing the flip of tyrosine residue in aspart (magenta cartoon, B26Tyr and B28Asp shown as sticks), when compared to the glycine intermediate (green cartoon, B26Tyr and B28Gly shown as sticks) b) No flip of the tyrosine residue seen in WT-is (cyan cartoon, B26Tyr and B28Pro shown as sticks). This is again compared to the glycine intermediate (brown cartoon, B26Tyr and B28Gly shown as sticks). The sequence change accompanying the transformation of WT-ins to Aspart is also shown.

Disease Mutants:

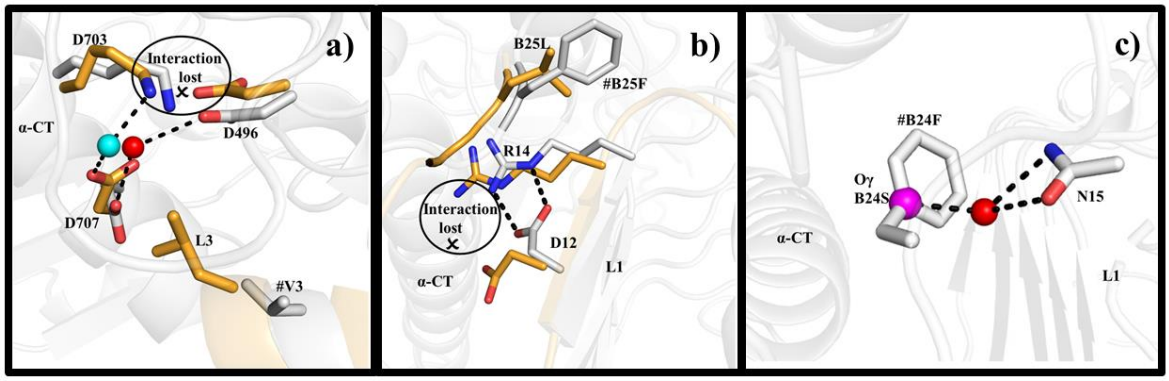

Figure 6: Structural changes accompanying disease-causing mutations are shown. Hydrogen bonds are indicated by dashed lines. In all figure panels WT-ins-IR complex is shown with carbons in grey and the corresponding disease-mutant-IR complex is depicted with carbons in orange. In general the complexes are represented as cartoons with key resides from the receptors and ligands highlighted as sticks. **a**) insulin Wakayama, showing the loss of a water-mediated interaction network between two domains in the IR. The red sphere represents the bridging water in the WT-ins-IR complex between D496, D707 and K703. The cyan sphere shows the displacement of this bridging water due to the larger Leucine residue in A3, which causes the loss of the inter-domain interaction with D496 in this complex; **b**) insulin Chicago the salt-bridge between D12 and R14 within the L1 domain is present in the WT-ins-IR complex and is lost upon mutating B25Tyr to Leucine, and **c**) Insulin Los Angeles depicts the presence of B24Phe in a buried pocket within the IR. Mutation to the smaller serine resulted in partial solvation of this pocket, where the residue was observed to forms a water-mediated link to the N15 residue within the L1 domain.

B10 Substitutions:

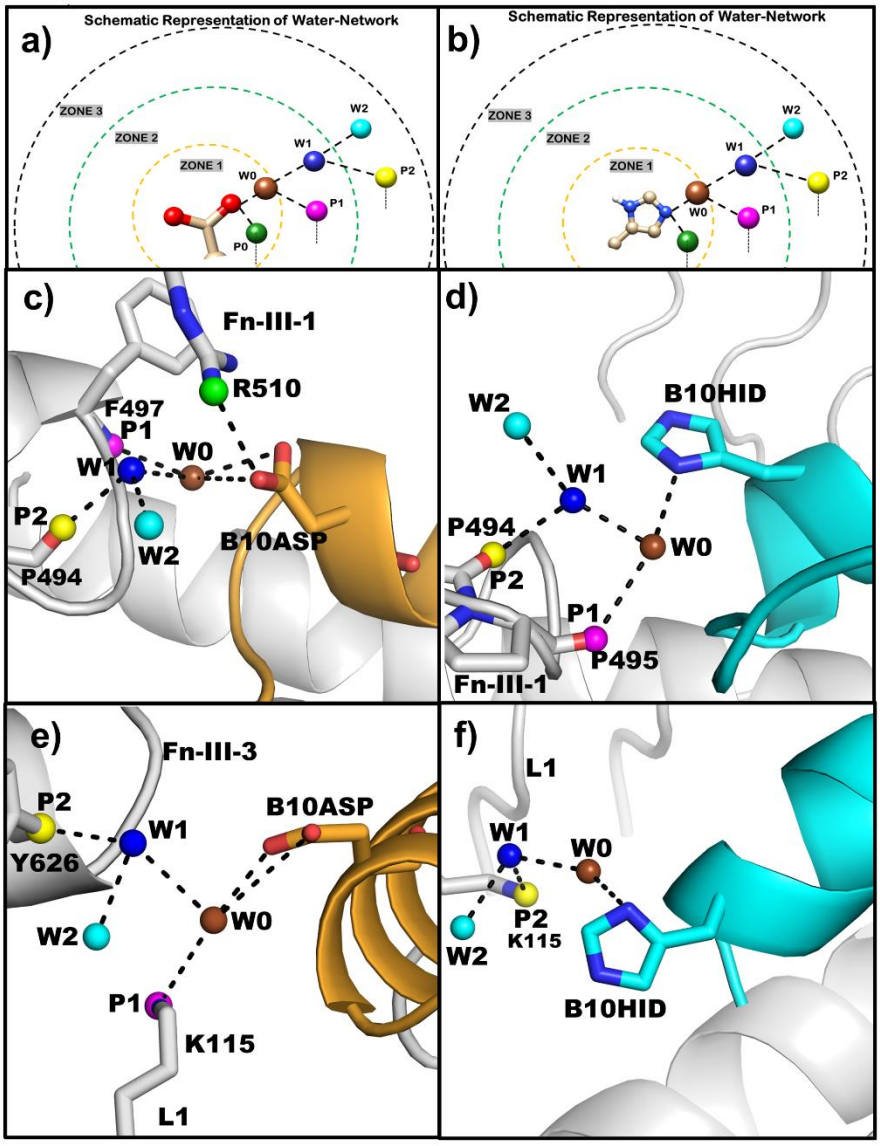

Figure 7: Analysis of the Water-mediated interaction network around the position B10 of insulin in both IR and IGF-1R. **a-b**) A schematic representations of the classification of water-mediated interactions around a residue is shown. The Aspartic acid or histidine residues occupying position B10 are shown in ball & stick representation, polar oxygen and nitrogen atoms are shown red or blue spheres respectively. First order or direct polar interactions with between B10 and a receptor residue is considered PO and similarly W0 for a water contact. Second order contacts are classified as a WO-receptor polar contact defined as P1 or W0-water contact defined as W1. Similarly, third order contacts are defined as W1-receptor polar contact

defined as P2 or W1-water contacts defined as W2. P0-2 and W0-2 are color coded as shown in the figure. **c-f**) Shows the water mediated contacts for B10Asp-IR (c), B10His-IR (d), B10Asp-IGF-1R (e) and B10His-IGF-1R (f). The figures highlight the increase in water-mediated polar interaction networks, when B10 is occupied by an Aspartic acid residue when compared to B10His in both the IR and IGF-1R.

Truncated Insulins:

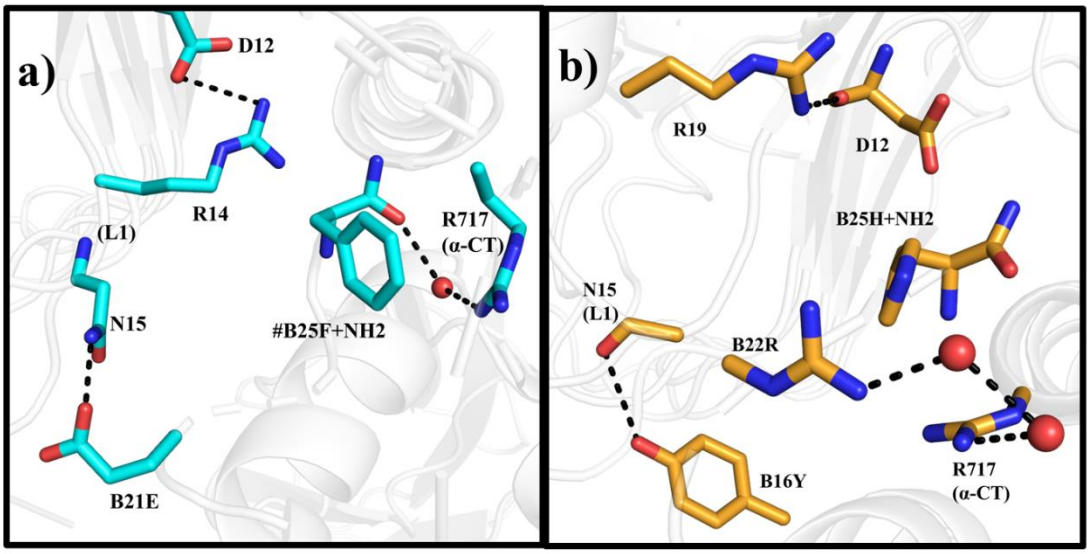

Figure 8: Structural analysis of truncated insulins (desB26-B30). In a) Truncated insulin with C-terminus occupied by the natural WT residue i.e. B25Phe is shown. In general the receptor-ligand complex is shown as a cartoon with carbon atoms in grey. The highlighted interaction networks are shown with participating residues in sticks with carbon atoms in cyan b) Truncated insulin with C-terminus occupied by the mutant B25His is shown. In general the receptor-ligand complex is shown as a cartoon with carbon atoms in grey. The highlighted interaction networks are shown with participating residues in sticks with carbon atoms in orange.

Supplementary Information for

Mapping the Structural Drivers of Insulin Recognition and Specificity Using Molecular Dynamics and Free Energy Calculations

Table S1**:** A summary of currently available XRD and Cryo-EM structures for IR and IGF-1R receptors in complex with different ligands or as apo receptors.

Receptor		**Method**	**PDB IDs**	**#SS**	**Year**	**Binding Site**	**Ref.**
IR	High affinity insulin analogs	XRD	3W12	3	2013	HA	(4)
			3W13	3	2013	HA	(4)
			5KQV(3W14)				
IR	Insulin (wild type)	XRD	4OGA	3	2014	HA	(20)
IR/IGF-1R Hybrid	IGF-1	XRD	4XSS	3	2015	-	(33)
IR-apo	-	XRD	4ZXB	-	2016	-	(34)
IR	Insulin (wild type)	EM	6CE9	1	2018	HA	(2)
			6CEB	1	2018	HA	(2)
			6PXV	3	2019	Both	(16)
			6PXW	3	2019	Both	(16)
			6SOF	1/3	2020	HA/LA	(3)
			7QID	3	2020	HA/LA	(3)
			7PG4	2	2022	HA/LA	(17)
IGF-1R	IGF-1	XRD	5U8Q	3	2018	-	(11)
			5U8R		2018		(11)
IGF-1R	IGF-1	EM	6PYH	3	2019	HA	(40)
IGF-1R	Insulin (wild type)	EM	6JK8	3	2020	-	(68)

Note**:** #SS: Number of disulfide bridges in Ligand. BS: binding site type. HA: High affinity and LA: Low affinity

Table S2**:** Binding enthalpies calculated from MM-GBSA calculation. Mean, SEM and Sd. are calculated from the ΔH calculated from ligand molecule bound at two HA sites, s1 & s1' from two simulations with different initial velocities.

Complex	ΔH_{bind} (kcal/mol) (MM-GBSA)		SASA(Å^2) (LCPO, surf)	
	Mean	SEM	Mean	Sd.
6ce9-wtIns	-66.29	8.28	618.70	74.96
5u8q-Igf	-71.46	1.67	678.39	65.50
5u8q-wtIns	-55.06	7.31	643.88	60.55

Table S3: Calculated relative free energies ($\Delta\Delta G_{bind}$) at IR and IGF-1R calculated for all the insulin analogs considered in this study along with the range of experimental values available for each analog. The experimental potency values are shown as an inner-range (spread of means) and outer-range (absolute spread including reported error ranges). The experimental spread of $\Delta\Delta G_{bind}$ has been calculated using the inner means and a mean of means has been considered in order to compare with the calculated $\Delta\Delta G_{bind}$. The experimental and calculated $\Delta\Delta G_{bind}$ have both been reported in relation to WT-Ins or equivalent in the case of truncated systems.

Molecule Descriptor	$\Delta\Delta G_{bind} \rightarrow$ WT-Ins (kcal/mol)		Mean $\Delta\Delta G_{bind}$ from experimental potencies (inner range of values shown). kcal/mol		Experimental Potencies relative to WT-ins (%). WT-ins set to 100%.		Deviation from the experimental mean value [$\Delta\Delta G_{bind}$ (exp) - $\Delta\Delta G_{bind}$ (FEP)] (kcal/mol)	
	IR	**IGF-1R**	**Metabolic**	**Mitogenic**	**Metabolic**	**Mitogenic**	**Metabolic**	**Mitogenic**
Lispro	-0.43✓	-2.55▪✓	-0.08 (-0.14 to 0.01)	-0.18 (-0.3 to 0.01)	79-85(79-99) (45,69)	76-79(56-99)(45,69)	+0.35 (Dark Green)	+2.3 (Orange)
Aspart	0.30✓	-4.49▪✓	0.01 (-0.01 to0.02)	-0.21 (-0.01 to0.02)	99-103 (45,69)	36-80(33-150) (45,69)	+0.29 (Dark Green)	+4.2 (Orange)
A21Gly	-0.71✓	+0.47×	-0.08 (-0.06 to -0.1)	-0.5 (-0.06 to -0.1)	85-91 (45,70)	46-64 (22-90) (45,70,48)	+0.63 (Dark Green)	-0.97 (Red)
Wakayama	-0.69○✓	NA	-3.20 (-3.0 to-3.4)		0.4-0.8 (23)		-2.5 (Orange)	
Chicago	-0.79○✓	NA	-2.37 (-2.3 to-2.5)		1.9-2.5 (23)		-1.5 (Orange)	
Los Angeles	+0.55×	NA	-2.15 (-3.2 to-1.1)		0.6-16 (23,26)		**-2.7** (Red)	
B10D	+1.64▪✓	+2.98▪✓	0.47 (0.4-to 0.5)	1.1 (0.8 to 1.5)	195-221 (193-237) (27,45)	733-802(403-1148) (27,45)	+1.1 (Orange)	+1.7 (Orange)
B10A	-1.05✓	-0.73✓	-0.19 (-0.3 to-0.1)	-0.88 (-0.7 to 1.0)	62-86 (27)	18-32 (27)	+0.8 (Dark Green)	-0.16 (Dark Green)
B10Q	+3.3✓▪	+0.35✓	0.01(-0.1 to 0.12)	0.33 (0.14 to 0.85)	83-123 (27)	126-234 (27)	+3.2 (Orange)	+0.01 (Dark Green)
B10V▲	-0.3✓	-0.01✓	-0.26 (-0.3 to-0.1)	-0.34(-0.6 to -0.1)	54-80 (27)	35-95 (27)	0.03 (Dark Green)	-0.33 (Dark Green)
B25H*	1.53✓	NA	0.7 (0.6 to0.8)		261-375 (32)		+0.8 (Dark Green)	
B25Y*	1.99▪✓	NA	0.57 (0.4 to0.7)		192-330 (32,30 ,30)		+1.4 (Orange)	
B25W*	+2.1×	NA	-0.38 (-0.5to 0.3)		44-66 (32)		+2.5 (Red)	

✓ Indicates the direction of shift predicted from FEP is in-line with experiments.

× Indicates the direction of shift predicted from FEP is not in-line with experiments

○ Indicates the magnitude of shift predicted from FEP is under estimated from the mean with a difference in magnitude of >1 kcal/mol

■ Indicates the magnitude of shift predicted from FEP is over estimated from the mean with a difference in magnitude of > 1 kcal/mol

∗ Indicates systems with (des B26-B30) and B25 position is amidated

⅄: Fep calculations were performed using desB30 system.

The color below the deviation from experimental mean values indicate the accuracy of the prediction. Dark green indicate that the direction of shift was predicted correctly and magnitude of difference between FEP and mean experimental values are < 1 kcal/mol. Orange indicates that the direction of shift predicted accurately and magnitude of difference between FEP and mean experimental values are > 1 kcal/mol. Red indicates the direction of shift predicted inaccurately by FEP. The same color scheme is followed in Figure 4 of the main text.

Table S4: Ensemble averaged solvent interaction networks for B10Asp & B10His in IR/IGF-1R from MD trajectories.

	Protein		Water		Protein		Water	
	IR		IR		IGF-1R		IGF-1R	
	B10Asp	B10His	B10Asp	B10His	B10Asp	B10His	B10Asp	B10His
Zero order contacts	2.24 (P0)	0.06 (P0)	4.38 (W0)	2.84 (W0)	0 (P0)	0 (P0)	6.68 (W0)	2.74 (W0)
First order contacts	2.68 (P1)	0.76 (P1)	13.08 (W1)	9.16 (W1)	0.24 (P1)	0.08 (P1)	25.94 (W1)	9.98 (W1)
Second order contacts	28.16 (P2)	8.26 (P2)	40.10 (W2)	15.98 (W2)	2.02 (P2)	0.64 (P2)	38.84 (W2)	22.38 (W2)

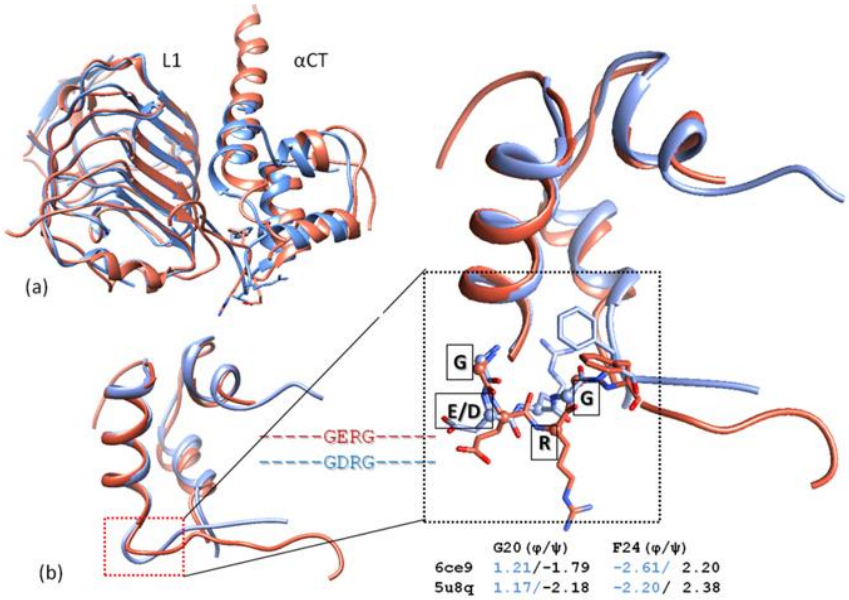

Figure S1: (a) Superposition of major ligand binding domains of the IR (Brown, PDB: 6CE9) and IGF-1R (Blue, PDB: 5U8Q). Bound conformations of WT-ins and IGF-1 with their cognate receptors, IR and IGF-1R, respectively are shown using the same color scheme. The "GXXG" motif and B24Phe are shown as ball and stick, and sticks, respectively and are highlighted in the figure. The dihedral angles calculated for hinge motif confirm the similarity in binding conformations of the ligands.

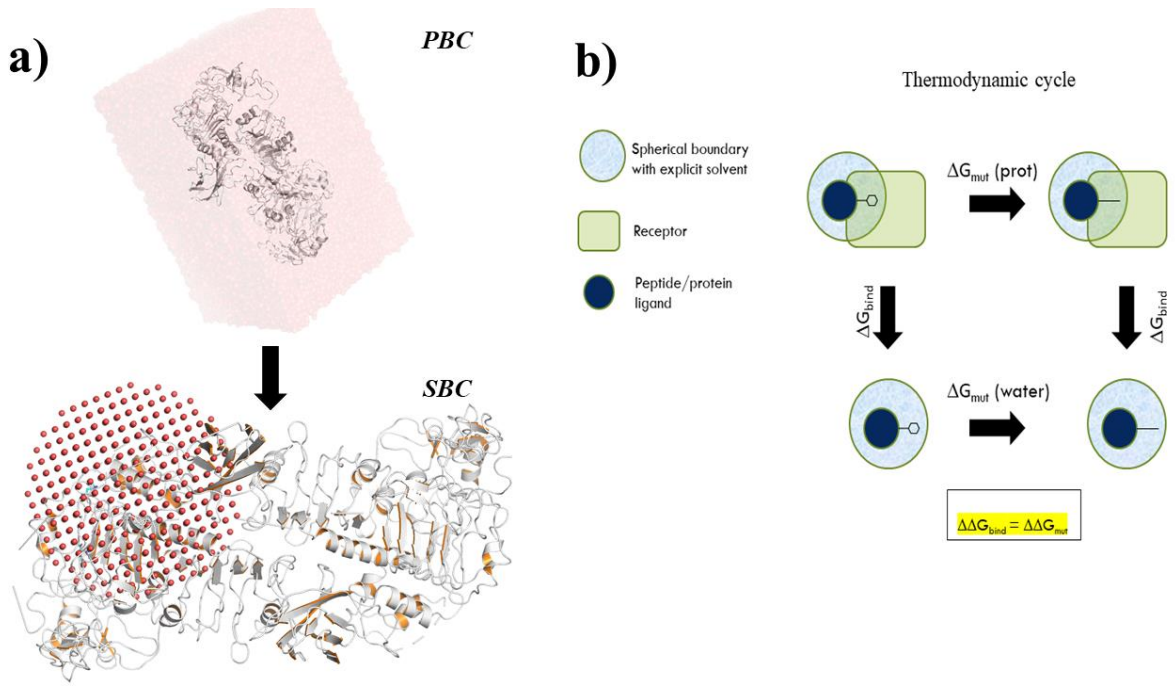

Figure S2: a) depicts periodic boundary conditions (PBC) used for MD simulations and spherical boundary conditions (SBC) used for MD-FEP calculations. The top panel shows water molecules around the protein (cartoon) as surface representation in the PBC simulations. The bottom panel shows SBC used for FEP calculations in this study. All atoms within the sphere are fully flexible whereas those outside the sphere are restrained to their initial coordinates. The water molecules are shown as red spheres with receptors shown as cartoons. b) Schematic representation of thermodynamic cycle used for FEP calculations. The blue circle represents the sphere, green rounded rectangle represents receptor and filled dark blue circle represents the peptide ligand. The same mutations are performed with the peptide bound to the receptor and with free peptide in water. The relative $\Delta\Delta G_{bind}$ is calculated by taking advantage of the thermodynamic cycle as shown.

Extended structural analysis of B10 mutants

B10A: FEP calculations accurately captured the loss of function both at IR & IGF-1R associated with this mutation which is in line with experiments (Table S3, Figure 4). In IR, there was negligible change in the electrostatics compared to the reference state with the major loss of favourable interactions accompanying the annihilation of the histidine side chain. Analysis of The structural changes upon the HisB10Ala mutation revealed a loss of contact between B13Glu and Tyr67 from the L1 domain and an inter-domain salt-bridge between R498 of Fn-III and E97 of L1 (Figure S3a). In IGF-1R, water mediated interactions were observed between His697 of the α-CT domain, Arg54 of FnIII-1 domain with residues B8 and B9 were disrupted upon mutation of B10His to alanine (Figure S3b).

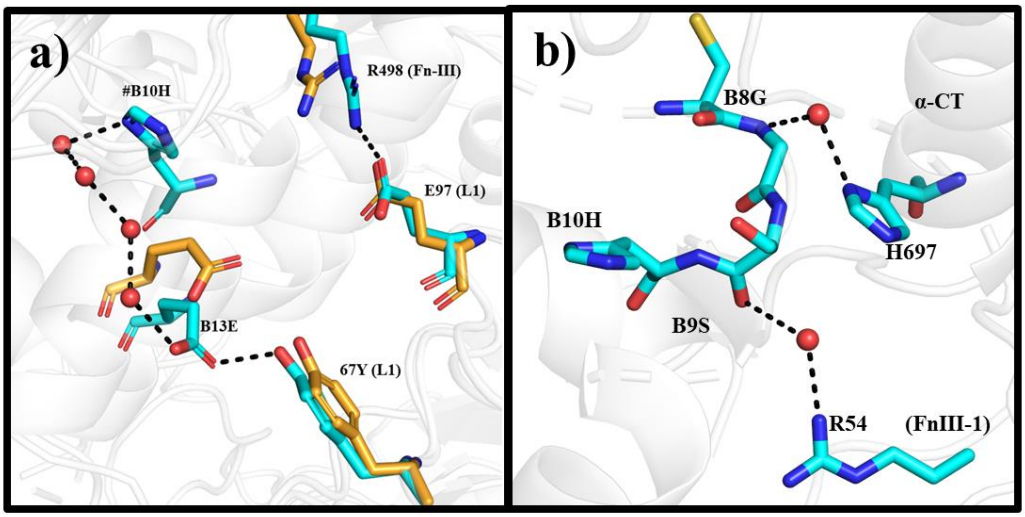

Figure S3: Structural changes of wildtype and B10A in a) IR & b) IGF-1R were shown. In a) the interacting residues in WT-ins and mutant B10A are shown as sticks with carbon atoms in cyan and orange respectively. Hydrogen bonds are shown as dashed lines. b) the water-mediated interactions of the in WT-ins within IGF-1R are shown with key resides as sticks with carbon atoms in cyan. The water oxygen atom is represented as a red sphere. The receptors are shown as cartoons with carbon atoms in white.