Article

Folding and Intrinsic Disorder of the Receptor Tyrosine Kinase KIT Insert Domain Seen by Conventional Molecular Dynamics Simulations

Julie Ledoux, Alain Trouvé [ID] and Luba Tchertanov *

Centre Borelli, CNRS, ENS Paris-Saclay, Université Paris-Saclay, 4 Avenue des Sciences,
F-91190 Gif-sur-Yvette, France; julie.ledoux@ens-paris-saclay.fr (J.L.); alain.trouve@ens-paris-saclay.fr (A.T.)
* Correspondence: Luba.Tchertanov@ens-paris-saclay.fr

Abstract: The kinase insert domain (KID) of RTK KIT is the key recruitment region for downstream signalling proteins. KID, studied by molecular dynamics simulations as a cleaved polypeptide and as a native domain fused to KIT, showed intrinsic disorder represented by a set of heterogeneous conformations. The accurate atomistic models showed that the helical fold of KID is mainly sequence dependent. However, the reduced fold of the native KID suggests that its folding is allosterically controlled by the kinase domain. The tertiary structure of KID represents a compact array of highly variable α- and 3_{10}-helices linked by flexible loops playing a principal role in the conformational diversity. The helically folded KID retains a collapsed globule-like shape due to non-covalent interactions associated in a ternary hydrophobic core. The free energy landscapes constructed from first principles—the size, the measure of the average distance between the conformations, the amount of helices and the solvent-accessible surface area—describe the KID disorder through a collection of minima (wells), providing a direct evaluation of conformational ensembles. We found that the cleaved KID simulated with restricted N- and C-ends better reproduces the native KID than the isolated polypeptide. We suggest that a cyclic, generic KID would be best suited for future studies of KID f post-transduction effects.

Keywords: receptor tyrosine kinase; RTK; KIT cytoplasmic region; kinase insert domain; KID; molecular dynamics and folding; conformational plasticity; intrinsically disordered region; transient structures; free energy landscape

1. Introduction

Receptor tyrosine kinases (RTKs) act as sensors for extracellular ligands, the binding of which trigger dimerisation of RT, activation of its kinase function and auto-phosphorylation of specific tyrosine residues in the cytoplasmic domain [1,2]. This leads to the recruitment and activation of multiple downstream signalling proteins, which carry the signal to the nucleus, where it alters patterns of gene transcription such as those governing various aspects of the cell physiology. Initiation of this cascade-like process involves different regions of the multidomain RTKs, each of them performing specific actions that are finely concerted by a tightly regulated allosteric mechanism controlling all functional processes of RTK [3]. Explicit elucidation of the signalling cascade represents an important and unsolved problem in cell biology.

The modular extracellular domain (ED) of RTK, containing characteristic structural motifs involved in specific ligand binding, is formed by various motifs (Ig-like, cysteine-rich, cadherin fragments, etc.) interconnected by coiled linkers, providing high conformational plasticity (Figure 1A). The binding of ligands affects the monomer–dimer equilibrium of RTKs by stabilising the dimeric state through a global conformational change in the ED [4]. The signal induced by ligand-binding propagates across the transmembrane (TM) domain to the cytoplasmic domain (CD) and promotes its activation. The cytoplasmic domain of

RTKs also has a modular structure composed of the juxtamembrane region (JMR), the split tyrosine kinase (TK) domain with the proximal (N-) and distal (C-) lobes linked by a kinase insert domain (KID), and the C-terminal tail (C-term).

We concentrate our study on the RTK KIT from the PDGFR (III-type) family. The physiological actions of KIT controlling cell survival, proliferation, differentiation, and migration depend on the activation of specific or overlapping pathways [5], which endows the activity of the SCF/KIT system (where SCF is the stem cell factor that regulates the KIT activation and therefore, triggers the initiation of multiple signal transduction pathways) of great complexity. Aberrant regulation of KIT signalling networks is associated with the progression of many cancer types, including human acute myeloid leukaemia, aggressive systemic mastocytosis, melanoma, gastrointestinal stromal tumour, and stomach cancers [6–8]. Disclosure of the KIT activated pathways in carcinogenesis will be a crucial step towards the development of KIT targeted therapies [9].

Functions of the TK domain of KIT, similarly to other RTKs, are mainly attributed to catalytic activity and trans-phosphorylation, while the post-transduction processes and the binding of intracellular proteins are associated with JMR, KID and C-terminal, which are the regions possessing multiple phosphorylation sites [1] (Figure 1B). The overall structural feature of these flexible fragments is the intrinsic disorder which is a vital condition for the creation of the dynamic networks of interacting proteins [10].
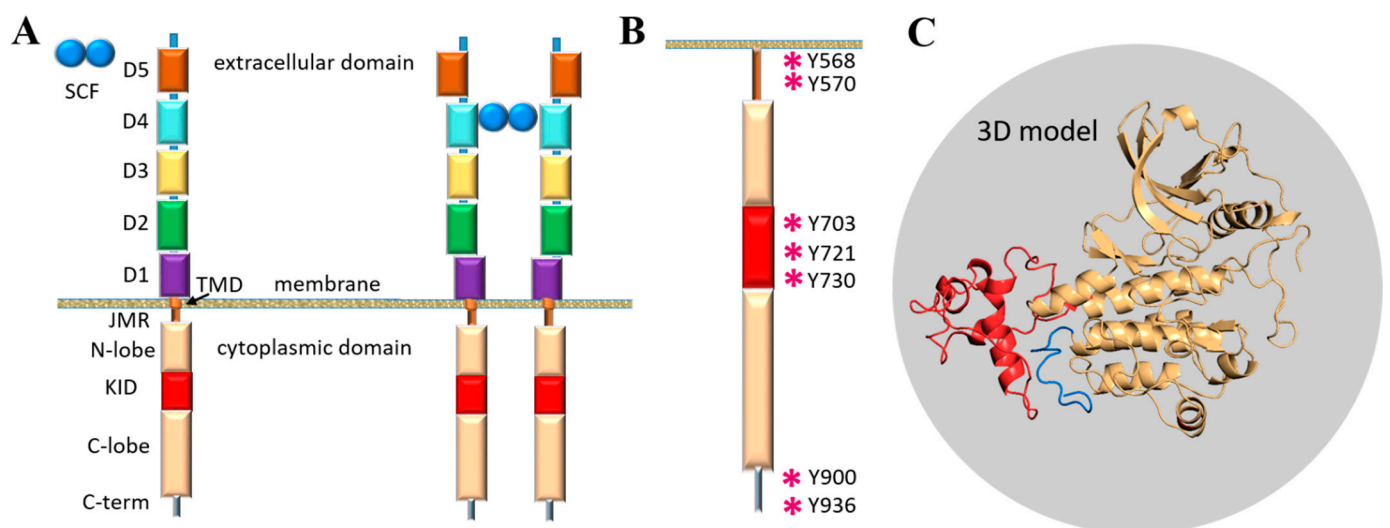
Figure 1. The modular structure of RTKs illustrated with KIT, a member of the RTK family III. (**A**) Structural composition of KIT: an extracellular domain (ECD) with five Ig-like regions (D1–D5), a transmembrane domain (TMD) and a cytoplasmic domain (CD) comprising a juxtamembrane region (JMR), an ATP-binding region (N-lobe), the phosphotransferase domain (C-lobe) with a kinase insert domain (KID) and a C-terminal. The stem cell factor (SCF) extracellular binding induces dimerisation and activation of KIT. (**B**) The tyrosine residues (Y, showed by asterisk) of JMR, KID and C-terminal identified as the phosphorylation sites involved in the recognition of cellular partners. (**C**) 3D structural model of the KIT CD with KID (red) and C-term (blue) [11].

A first structural model of the full-length cytoplasmic domain of KIT has been published (Figure 1C) [11]. The primary analysis of the structural and dynamical properties of this model showed that the conformational variability of KIT is provided mainly by JMR, KID and C-terminal, which were interpreted rather as intrinsically disordered (ID) regions demonstrating significant structural and conformational plasticity.

Because ID proteins (IDPs) lack stable secondary structures (2D) under physiological conditions, they exist as heterogeneous conformational (3D) ensembles, and are capable of rapidly changing conformation upon influence by an effector (e.g., binding of ligand/cofactor/protein). Principally, the determination of a single structure of the ID regions or ID proteins has no physical relevance, as such structures present only an isolated element

(a lone conformation) from a huge conformational space. A more pertinent approach is the description of such proteins in terms of the probabilistic population of the different regions, and the correlation of these probabilities with the protein function. Most experimental techniques employed to study IDPs [12] suffer from a conundrum: the empirical observables, which make it possible to assess the protein disorder, represent an average over the conformational ensemble, but the ensemble itself cannot be unequivocally inferred from the experiments. Therefore, computational methods provide an advantageous approach for analysing IDPs. In particular, molecular dynamics (MD) simulation provides great insights into this challenge [13], as an approach that can, in principle, produce the structural ensemble of biomolecules.

In the present study, the focus is on the KID of KIT, a binding hub assuming exquisite specificity of the receptor. As has been reported that five functional phosphorylation sites of KID from KIT, three tyrosine (Y703, Y721, Y730), and two serine (S741 and S746), provide the alternative binding sites for the adaptors, signalling and scaffolding proteins in the cytoplasm [14]. Phosphorylation of Y703 supplies the binding site for the SH2 domain of Grb2, an adaptor protein initiating the Ras/MAP kinase signalling pathway. Phosphorylated Y721 and Y730 are the recognition sites of PI3K and phospholipase C (PLCγ), respectively. The function of Y747 has not yet been described. Phosphorylated serine residues, S741 and S746, bind PKC (protein kinase C) and contribute to re-control of PKC activity under the receptor stimulation. The functional importance of KID is also emphasised by newly identified mutations of K704, N705 and S725, which were reported as activating in gastrointestinal stromal tumours [15]. Consequently, a study of KID structure-dynamics features and their relation to KID function is still crucial but/and obviously challenging.

The conformational dynamics of KID from KIT was probed by conventional molecular dynamics (cMD) simulations. This method generates the atomic-level data required for a detailed analysis of the conformational space and the identification of folding intermediates related to function and/or characterisation of functionally important phenomena related to allosteric regulation [16]. Allostery is often a fine adaptive mechanism of protein modulation in the cellular environment (e.g., adaptation of binding partners to form biomolecular assemblies) during signal transduction, catalysis, and gene regulation [17,18].

One approach for studying the assembly of multidomain proteins and their folding is to use the modular domain of the protein, which preserves binding capabilities even when the domain is removed from the context of the full-length protein [19]. The ability of modular protein domains to independently fold and bind both in vivo and in vitro has been taken advantage of by a significant portion of proteomics studies that have used modular domains to assess the protein–protein interactions required for a diverse set of cellular processes, including signal transduction and subcellular localisation.

The use of KID as a cleaved polypeptide represents a promising strategy for the exploration (empirically and numerically) of KIT signal transduction and the modulation of protein function by controlled interference with the underlying molecular interactions. Such use will be fully justified if we can prove that KID of KIT is a modular domain that preserves its structural and dynamic properties. As we seek draw conclusions with respect to the usefulness or the disadvantage of using cleaved KID as a reduced model to study the RTK KIT 'interactome' [20], a comparative analysis of this model with the KID of KIT was carried out. We also suggest a possible dependence of KID on the kinase domain of KIT, which can either be induced locally by geometric restriction on the KID terminus or globally promoted by long-range allosteric effects.

Our present study of KID from KIT focuses on its folding features and intrinsic disorder with the aim of defining a KID species suitable for the exploration of the KIT 'interactome'.

2. Results

2.1. Data Generation

The kinase insert domain of KIT, composed of the 80 amino acids (aas) (F689–D768), was examined by conventional MD simulations (all-atom, with explicit water) as the cleaved polypeptide (KID^C) and as the subdomain of the kinase domain (KID^D from KIT). The cleaved KID was simulated as a fully unconstrained entity (KID^C), and as a polypeptide with restrained distance between two Cα-atoms of terminal residues F689 and D768 (cleaved restrained KID, KID^{CR}), which is conserved in crystallographic structures of KIT [11]. The MD simulation run (of 1.8 μs for KID^C and 2 μs for KIT and KID^{CR}) was repeated four (KID^C) and two (KIT and KID^{CR}) times with different randomised initial atomic velocities to extend conformational sampling of each studied entity and to examine the consistency and completeness of the produced KID conformations. The other parameters of the MD simulations of KID and KIT were strictly identical. Each simulation was started with an equilibrated structure obtained after minimising the neutralised solvated protein which is either a 3D de novo model of KID or of KIT.

2.2. General Characterisation of MD Simulation Data

The data analysis was performed on each trajectory after the least-square fitting of MD conformations to the initial KID conformation, to avoid the motion of the domain as a rigid body.

First, the global stability of KID throughout the MD simulations was estimated using the root mean square deviations (RMSDs) computed on the Cα atoms relative to the initial KID model (at t = 0 μs) as a frame of reference. For the cleaved KID, simulated as the unconstrained polypeptide (KID^C), the RMSD profiles differ among trajectories 1–4, and their values vary along each trajectory, but the ranges of RMSD variations are comparable between trajectories (Figure 2A, top panel). Well-resolved slopes on RMSD curves are either a single event showing a rapid increase of RMSD values, or a two-event process showing an alternating increase/decrease in RMSD. Such sudden changes in RMSD, observed at t = 0.40–0.55, 0.77–1.79 and 1.55–1.60 μs in trajectories 1, 2 and 4, respectively, reflect regular conformational transitions in KID suggested in [11] based on a single simulation. The RMSDs cover a large range of values (4–9 Å) arranged in two main peaks per trajectory with different (most trajectories) or comparable (one trajectory) populations (Figure 2B, top panel). Two main peaks in each replica are separated by 1–3 Å, while the most populated peaks, composed of similar conformations, from replicas 1–4 are shifted only by ~1 Å. While RMSD of MD conformations were calculated from the initial coordinates, which are identical in all trajectories, we can suggest (1) at least two groups of highly different KID conformations within each trajectory; (2) partial overlapping of conformations from different trajectories.

Similar to the cleaved KID, the RMSD curves of KID^D show either a rather dramatic increase or a series of small reversal changes (Figure 2A, mid-panel). The probability distribution of such RMSD variations appears as two distinct well-separated peaks or two overlapped distributions in a slightly narrow range with respect to KID^C (Figure 2B, mid-panel). RMSD of the cleaved KID simulated with the restrained distance between its N- and C-end vary within a range similar to KID^D; nevertheless, the probability distribution shows a single peak for each replica (Figure 2A,B, low panel).

Second, the root mean square fluctuations (RMSFs) were calculated for each residue of KID. As expected, the highest RMSFs are mainly observed for residues at KID^C extremities, while the fluctuations of these residues in KID^D and KID^{CR} are very limited. Examination of inter-residual contacts in KID^D led to the identification of two stable (occurrence of 97–99%) hydrogen bonds N–H···O liking two pairs of residues, C691–L764, and L766–F689, which maintain the N- and C-end of KID in very proximal position (Figure 2, right; Figure S1). Apparently, these strong H-bonds restrain the mobility of KID extremities as reflected in small RMSFs. Similar to KID^D, the N- and C-ends of cleaved KID are stabilised by H-bonds, but the other pairs of residues are involved.
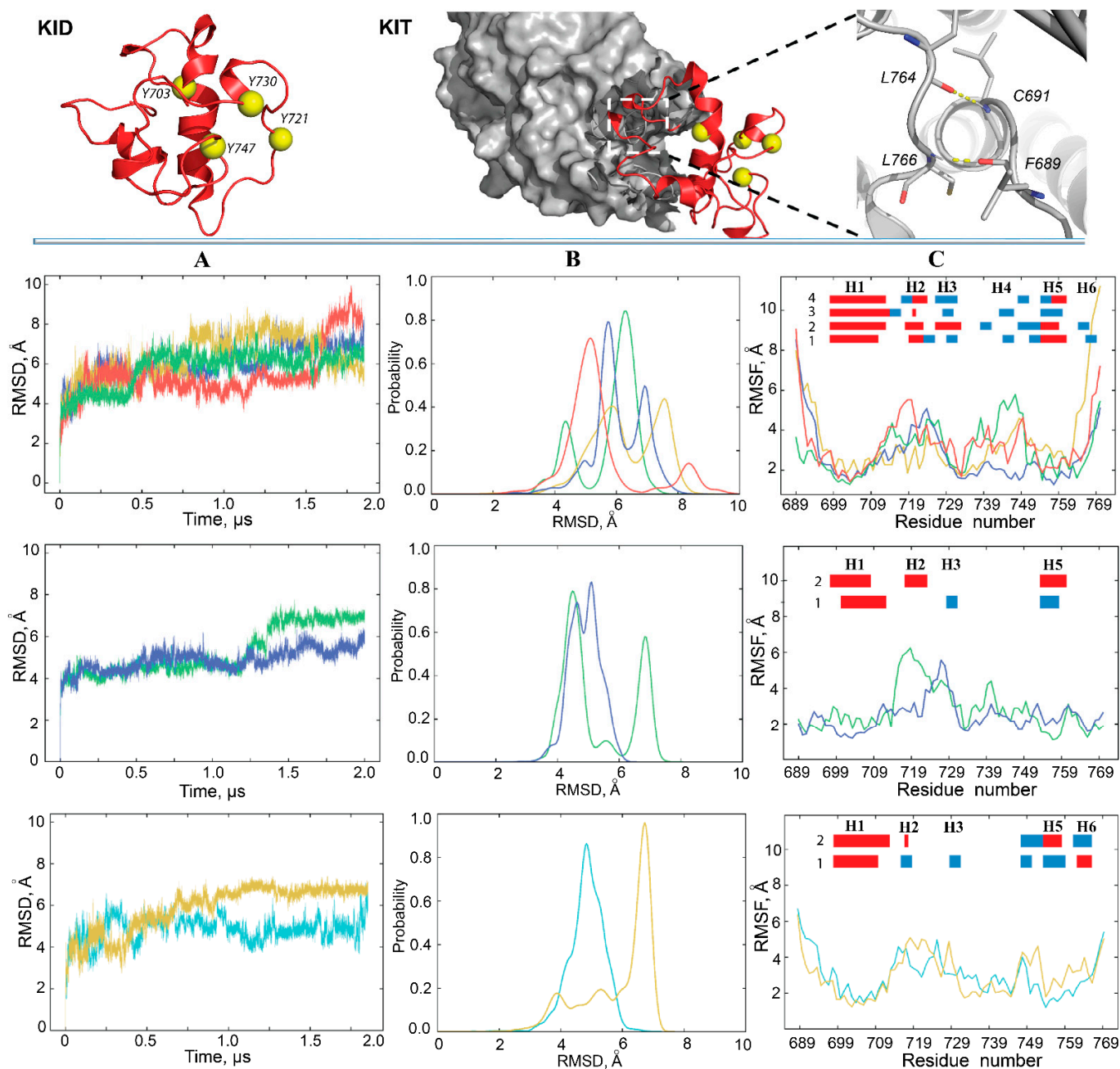
Figure 2. Conventional MD simulations of KID. (Top panel) Structural models of the cleaved KID (KID[C]) (left) and KID fused to KIT (KID[D]) (middle) in which the N- and C-ends are stabilised by pairs of H-bonds (right). Proteins are displayed as ribbons (KID), surface-filled model (kinase domain), yellow balls (the tyrosine residues) and sticks (residues formed H-bonds). Columns (**A**–**C**) show the statistical descriptors of KID[C] (top), KID[D] (middle) and KID[CR] (bottom) computed on the all Cα-atoms of KID for MD conformations of each trajectory after fitting on initial conformation. (**A**) RMSDs, (**B**) probability distributions of the RMSDs, and (**C**) RMSFs computed on the Cα atoms after fitting on initial conformation. In the insert, the folded secondary structures, αH- (red) and 3_{10}-helices (blue), labelled as H1–H6, were assigned for an average conformation of each MD trajectory. In (**A**–**C**), MD replicas are distinguished by colour: 1–4 of KID[C] are in green, yellow, blue, and red; 1–2 of KID[D] are in green and blue; 1–2 of KID[CR] are in blue light and yellow respectively.

Curiously, in all KID entities three hydrophobic residues, V731, V732 and P733, positioned on the random coil show small RMSFs in respect to preceded and followed residues, therefore, the RMSF curves in the different entities of KID display a comparable profile, which is described as the 'camel double-humped' contour (KID[C]) or close to this contour (KID[D] and KID[CR]) (Figure 2C).

The two descriptors, RMSD and RMSF, indicate the highly heterogeneous conformational composition of each data set obtained by cMD simulation of each KID entity. This data reveals the intrinsically disordered nature of KID from KIT, previously suggested in [11] based on a single trajectory of MD simulation. It is well known that a high content of polar and charged residues increases the propensity of a protein to be disordered [21–23]. The KID of KIT, composed of 58% such residues, is a good candidate for being an intrinsically disordered region, and this feature is probably only sequence-dependent and disconnected from the context of KID as an entity sampled by MD simulation, either as a cleaved isolated polypeptide or a domain of KIT.

Heterogeneous KID conformations were analysed with the ensemble-based clustering using the RMSD criterion [24]. To better grasp the conformational diversity of each studied KID entity, the conformations were clustered with the RMSD threshold r = 3, 4 and 5 Å. With r of 4 Å, we obtained a reasonably limited number of clusters regrouping almost all the conformations generated on each trajectory (Figures S2–S4). Since some clusters' representative conformations of different replicas showed striking similarity, we suggest that the conformational spaces generated by the independent trajectories of each KID entity partially overlap. Clustering analysis of the merged (concatenated) trajectory with the same RMSD threshold (r = 4 Å) produced the number of clusters which is less than a sum of clusters obtained for each replica. Such results do not contradict the supposed overlap of conformational spaces sampled by replicated trajectories.

2.3. Folding and Compactness of KID

The secondary structure interpretation (DSSP) indicates that the helical fold of KID is constituted of α- and 3_{10}-helices (Figure 2C; Figure S5). These helices are varied in number, length and $\alpha/3_{10}$ ratio. These variations are observed within a trajectory, between the trajectories, and for different KID species, but the position of some helices is curiously conserved.

KID^C is composed of six helices, H1–H6i is made up of 44–56% of amino acids. H1-helix, the largest (15–16 aas) of all KID helices, is a long-leaved α-helix well-conserved in all replicas. Other helices of varying lengths are rather transient, switching between α- and 3_{10}-helices (H2, H3 and H5) or between 3_{10}-helix and random coil, which is partially folded as reverse turn or bend (H4 and H6). Similarly, KID^D shows a helical fold, but the number of helices (H1, H2 and H3 or H5 on average structure) and the portion of residues (25–30%) constituting these helices are significantly diminished. In KID^D, like KID^C, H1-helix is the largest compared to other helices; however, its length is reduced to 11–12 aas. The H1-helix is long-lived, while the other helices are fully transient.

The highly diminished folding in KID^D prompts two hypotheses:

Hypothesis 1. *The KID folding depends on the KID ends, which are stabilised by strong H-bonds in KID fused to KIT and highly flexible in the cleaved KID.*

Hypothesis 2. *The KID folding depends on the status of KID as an entity, which is either autonomous polypeptide or collateral subdomain influenced by the kinase domain.*

The cleaved KID, simulated with the constrained distance between the terminal residues (KID^{CR}) showed that 30–35% of amino acids are involved in regular structures, the extended and long-lived α-helix H1 (15–17 aas) and three to five transient helices, indicating that the KID^{CR} folding (at the 2D level) is more ordered than in KID^D and less than in KID^C (Figure S5). It seems that the constrained polypeptide in general better represents the folding of the native KID than the unconstrained, but the expected equivalence was not observed. The mid-domain of KID is organised similarly in KID^{CR} and KID^D, while folding of C-terminal residues is quasi-identical in KID^{CR} and KID^C. It is probable that the applied constraints in KID^{CR} were rather soft compared to the restricted intra-protein geometry in KID^D. The increased flexibility of border residues in KID^{CR} compared to KID^D, as seen by RMSFs, and the greater similarity of 2D folding of residues from C-end (H6) in KID^{CR}

and KID^D, support this hypothesis. On the other hand, we can suggest a certain bias of such comparison derived from a different number of independent MD trajectories that is 4 for KID^C and 2 for KID^D and KID^{CR}. Interestingly, in all studied KID entities, residues forming H1 and H5 helices show the smallest RMSF values compared to the other helices.

The three-dimensional (3D) structure of KID in all studied entities represents a compact array of αH- and 3$_{10}$-helices linked by short or extended loops (random coils or turns) that play a principal role in the conformational diversity of KID. In the most of conformations, KID retains its collapsed globule-like shape that is slightly elongated towards the KID connected to the kinase domain, and therefore, is best described as a flattened (oblate) ellipsoid with an opening given by the distance between its ends (Figure 2). This opening in KID^D is controlled by hydrogen bonds linking residues from N- and C-extremities. The enlarged displacement of the border residues in the cleaved KID promotes a diminishing or a full destabilisation of such H-bonding.

To characterise the size and compactness of KID, the mass-weighted radius of gyration (*Rg*) was calculated for each entity of KID on MD conformations. We compared the distribution of *Rg* vs. RMSD between each ensemble of KID conformations (on every trajectory on each KID entity) (Figure S6). All distributions show at least two heavily populated regions separated by an area with a lower number of intermediates. The mean values of *Rg* for each peak are comparable in the two replicas on KID^D (and on KID^{CR}), but are slightly different between entities. In KID^D, the *Rg* of each peak (mv of 11.80(1) and 12.22(1) Å) shows a slightly increased value in comparison with KID^{CR} (mv of 11.33(1) and 12.00(2) Å). In KID^C, the *Rg* mean values of the heavily populated regions vary between replicas, but these variations spanned (cover) the *Rg* range observed in KID^D and KID^{CR} (mv of 11.64(1) and 12.52(2) Å).

Since strong variations in *Rg* (and RMSD) are observed along the same trajectory on cleaved KID, this indicates a higher conformational variability of KID^C, rather than more exhaustive sampling over the four replicates. Additionally, since the cluster analysis of the individual and concatenated trajectories showed at least a partial overlap of the conformational spaces generated on the replicas for each KID entity (Figures S2–S4), the more in-depth analysis of each KID entity was performed on the concatenated data.

2.4. Folding and Intra-Domain Interactions of KID

The propensity of globular proteins to be compact is the key reason that their folded states achieve high packing density. We suggest that the contact map can characterise KID collapsibility. The high instability of folded structures, which are, except for H1, transient, converting between αH- and 3$_{10}$-helices or between helix and random coil, together with their great mobility, producing an expected irregularity of contacts, resulted in the smeared pattern on the contacts maps (Figure 3A).

The contact maps are sufficiently different between replicas of the same entity of KID (e.g., replicas 1–4 of KID^C), but strikingly similar between distinct KID entities (e.g., replicas 3 on KID^C and 1 on KID^{CR}, or replicas 4 on KID^C and 2 on KID^{CR}). Such patterns in the contact map can be linked, on the one hand, to a large difference in folding of the same KID entity observed in the MD replicas, and on the other hand, to a partial similarity of the conformational and structural characteristics of the different entities of KID.

According to the contact maps, the residues of N- and C-ends of the cleaved KID, despite their high (in KID^C) or moderate (in KID^{CR}) flexibility, are still involved in intramolecular contacts. Apparently, the change in conformation in the cleaved KID leads to the appearance of alternative hydrogen bonds, which are formed by the same residues as in KID^D, and the other neighbour residues (C691 $\cdots$ D765, S692 $\cdots$ D765, L766 $\cdots$ F689, E767 $\cdots$ I690).
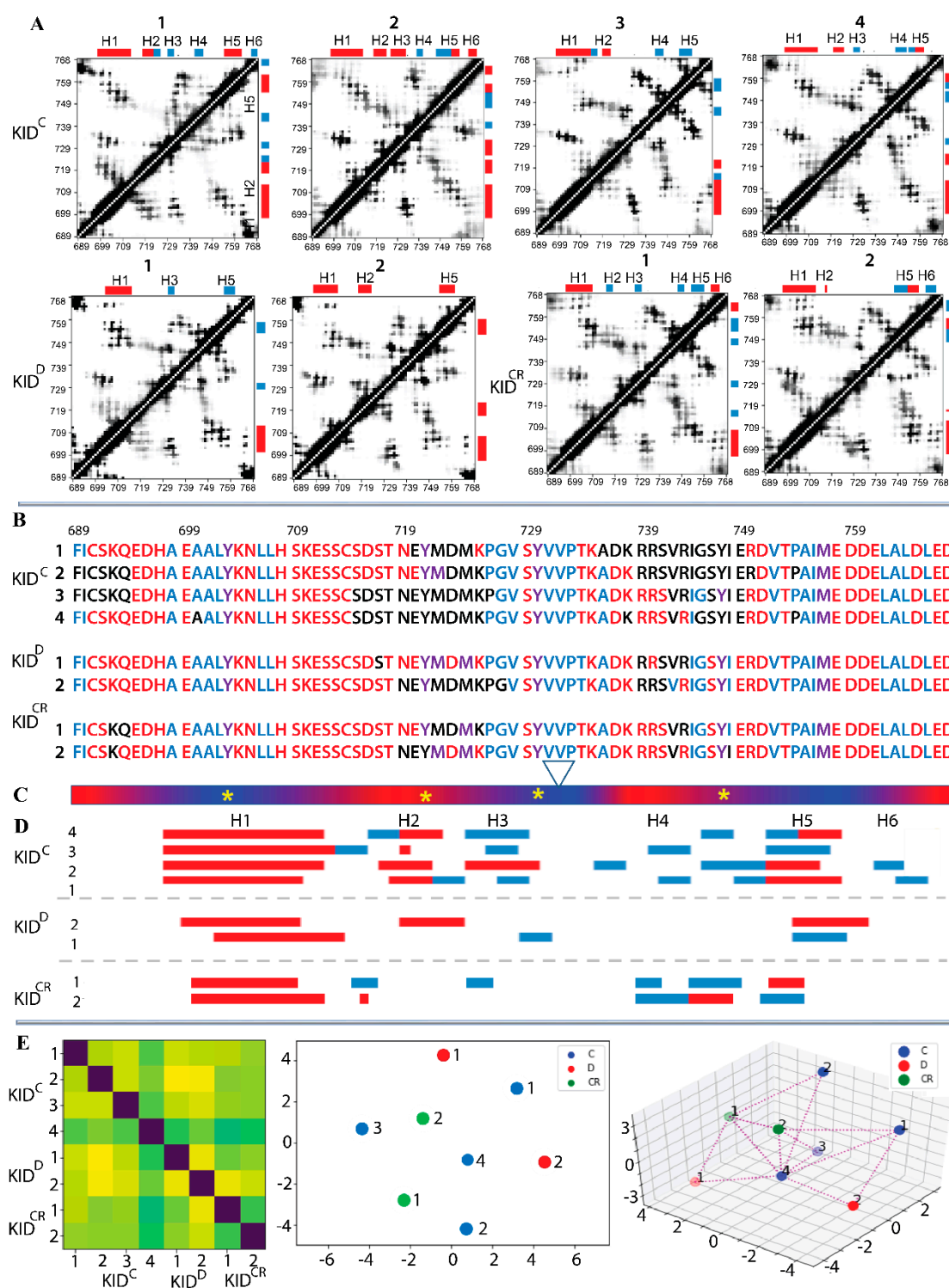
Figure 3. Estimation of intramolecular contacts in KID. (**A**) Dynamic contact maps of pairwise distances Cα−Cα < 10 Å computed for each MD trajectory. Black–white gradient shows a frequency (from 1 to 0) of contact during a trajectory. Secondary structure bars are shown at the top and on the left of the contact map. (**B**) Residues participating in steady contacts (in colour) were identified from (**A**) and analysed visually (PyMOL). Residues from the 'pseudo-rigid' segments E699–S709, P754–L764 and V731–P733 were used as an origin for the computing of contacts to all residues of KID. Only the contacts with an occurrence greater than 80% were considered. Polar, hydrophobic, and amphiphilic residues are denoted in red, blue and violet respectively. (**C**) Red–blue gradient shows the RMSF values, from large (>4 Å, in red) to small (<2 Å, in blue). The position of tyrosine residues is shown by the yellow asterisks. (**A**,**D**) The αH- (red) and 3_{10}-helices (blue), were assigned by DSSP on average conformation from each MD trajectory and labelled from H1 to H6. (**E**) The per-residue modelling

by finite-state Markov models of the secondary structure dynamics of KID. Transition probability matrix from one folding state to another (from one letter to another) was obtained on data encoding the secondary structure for each residue of each replica. (Left) The Fisher-Rao matrix (8×8 of size), where the first 4 replicas are the cleaved KID^C (group C), the next two the KID^D (group D), and the last two the cleaved restricted KID^{CR} (group CR). (Right) Multi-Dimensional Scaling (MDS) represents an 'as isometric as possible' embedding of the data in 2D and 3D (i.e., a representation by placing points on a plane and in Cartesian coordinates while preserving the calculated inter-distances as well as possible).

The high-occurrence of contacts between multiple residues from the folded and randomly coiled segments of the mid-domain of KID^C indicate diverse non-covalent interactions between helices (H1, H2 and H5), helices and coiled regions, and randomly coiled regions. This large number of intra-molecular contacts suggests that multiple regions contribute to maintaining the inherent (intrinsic) 3D structure of KID in all studied entities, independently of the fold.

While a considerable portion of KID undergoes both large conformational changes and high fluctuations, we focused on residues showing low fluctuations (RMSF)—the 'pseudo-rigid' segments E699-S709, P754-L764 and V731-P733—and analysed their contacts with all the other residues (Figure 3B,C). It is interesting to note that these 'pseudo-rigid' fragments show a different fold, a long-leaved helix (H1), a random coil linking H3 and H4 helices, and a transient helix (H5), respectively.

The most fascinating observations from the analysis of contacts formed by the 'pseudo-rigid' fragments are as follows: (i) a large number of KID residues are involved in intramolecular contacts, (ii) the charged and polar 'contacting' residues are the most abundant compared to hydrophobic residues, (iii) residues V728–D737 constituting the linker connecting H3 to H4 or H5 are regularly involved in intramolecular contacts, and (iv) residues of the N- and C-ends interact with the 'pseudo-rigid' segments. The quantitative estimation of the number of contacts stabilising KID is minimal in KID^D, while in the cleaved KID their number is increased to 46% (KID^C) and 52% (KID^{CR}). Interestingly, there are a fairly limited number of residues that do not contribute to intramolecular interactions in KIT.

Analysis of the contribution of 'contacting' residues in the intramolecular H-bonds (including the salt bridges) and the hydrophobic interactions, inspected separately, showed that both types of interactions form an extended and dense network of contacts that stabilises a compact globular shape of all studied KID entities (Figure 4A). To illustrate that these complex intramolecular contacts held a 'globule-like' shape of KID, we depicted the H-bonds and hydrophobic interactions on a randomly chosen conformation of each entity (Figure 4B).

In particular, the helices H1 and H5 interact mainly through the multiple hydrophobic interactions stabilising their proximal position. Residues from H1 and H5 form H-bonds and hydrophobic interactions with residues from the loop connecting H3 and H4, and/or from H4. These abundant interactions stabilise a closed location of the structurally disordered regions of KID mid-domain to helices H1 and H5. H1 interacts with H2 and H3 mainly via the H-bonds formed by charged and polar residues. Finally, the N- and C-end residues interact with each other and with H1-helix. The charged (E758, E595, E699, E711 and E761) and polar (N705, N719) residues form the H-bonds by the interaction of their side-chains with the main-chain atoms of the other residues.

The observed patterns of non-covalent intramolecular contacts in KID display a key role of the αH1-helix which, like a drop of glue, attaches all the structural fragments of KID around them. The results obtained indicate that intramolecular interactions, van der Waals and electrostatic forces, are a dominant factor in the stabilisation of the compact 'globule-shaped' (collapsed) KID and that the collapse induced by the intramolecular force conquers the solvent-induced expansion.
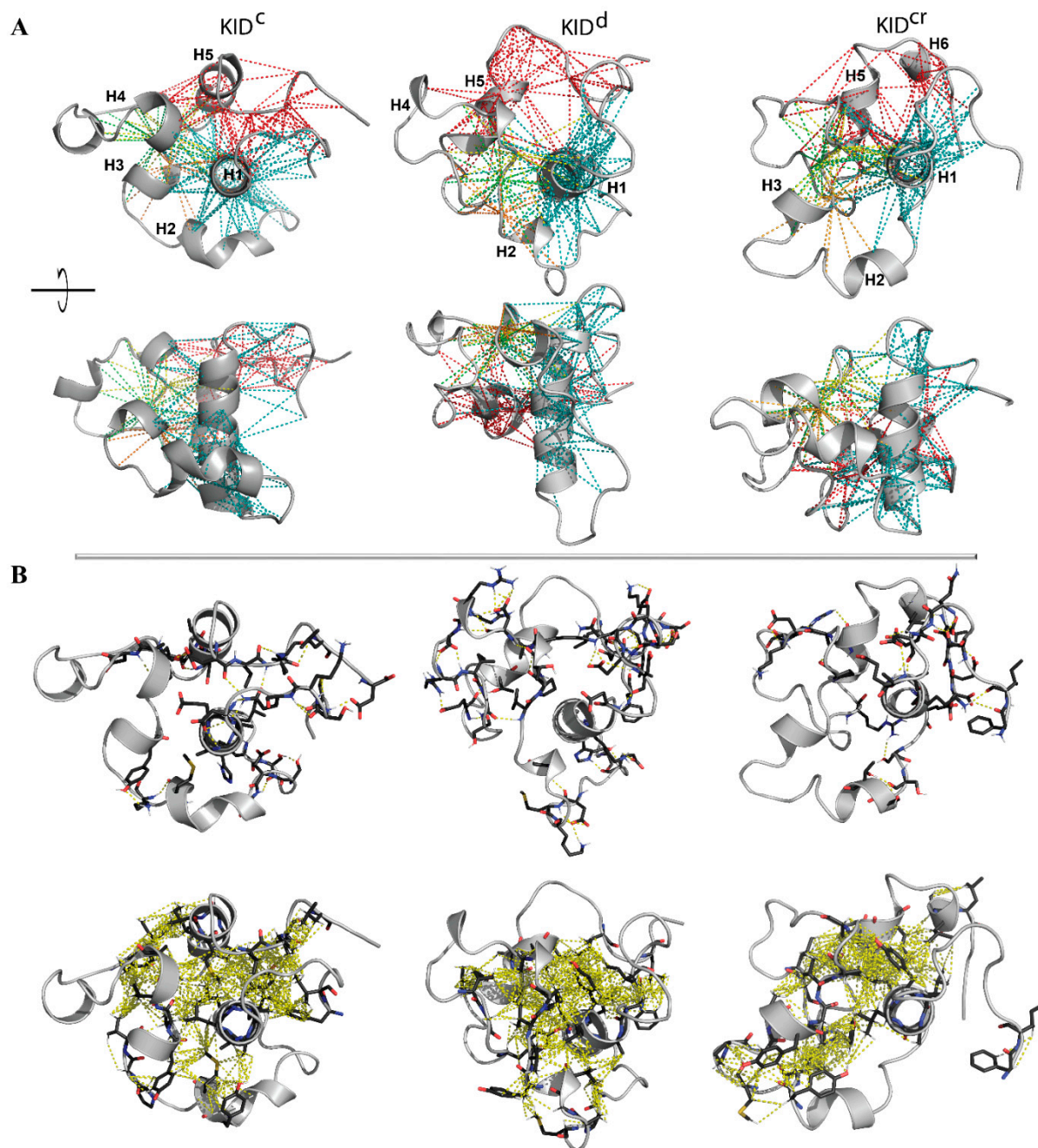
Figure 4. Non-covalent interactions maintaining the inherent (intrinsic) 3D structure of KID. (**A**) The intramolecular contacts, H-bonds and hydrophobic interactions (dashed lines), formed by residues E699–709 (H1) (in teal), P754-L764 (H5) (in red) and V731/V732/P733 (orange/green/yellow) with all the residues of KID at least in one trajectory, are superimposed on a randomly chosen KID conformation shown in two orthogonal projections. Protein is shown in grey ribbons, helices H1–H6 are labelled for KID^C. (**B**) H-bonds (D-H$\cdots$A $\leq$ 3.6 Å, where D (D = O/N/S) is a donor atom and A (A = O/N/S) is an acceptor atom) (top) and van der Waals contacts (C-H$\cdots$A = O/N/S $\leq$ 4 Å) between all side chain atoms (bottom) are shown on a randomly chosen conformation. (**A**,**B**) The interactions stabilising the regular structures (helices) are not considered.

Since the non-covalent contacts of KID are folding-dependent, we further focused on the time-connected dynamics of the KID secondary structure. To more finely compare the folding dynamics in KID entities, we carried out a study based on per-residue modelling by finite-state Markov models, which was carried out on data encoding the secondary structure for each KID residue of each replica in an estimated transition probability matrix

from one folding type to another. The eight-category classification (eight-letter code) of secondary structures was used. We thus obtain for each replica a sequence of transition matrices (one per residue). A suitable distance (Fisher-Rao distance) between these families is then calculated for all the pairs of replicas. We obtain the Fisher-Rao matrix (8×8 of size), where the first four replicas are the cleaved KID^C (group C), the next two the KID^D (group D), and the last two the cleaved restricted KID^{CR} (group CR) (Figure 3F). Multi-Dimensional Scaling (MDS) was performed to get an 'as isometric as possible' embedding of the data in 2D (i.e., a representation by placing points on a plane while preserving the calculated inter-distances as well as possible). Interestingly, the three groups do not form separate clusters, as shown in the 2D representation. Group C occupies the space quite well with replica number 4 at the central position, which is also central for all KID trajectories, while the two points of group D occupy extreme positions. These observations corroborate fairly well with the secondary structures (Figure 5C,E) and the inter-distance matrix (Figure 3A).

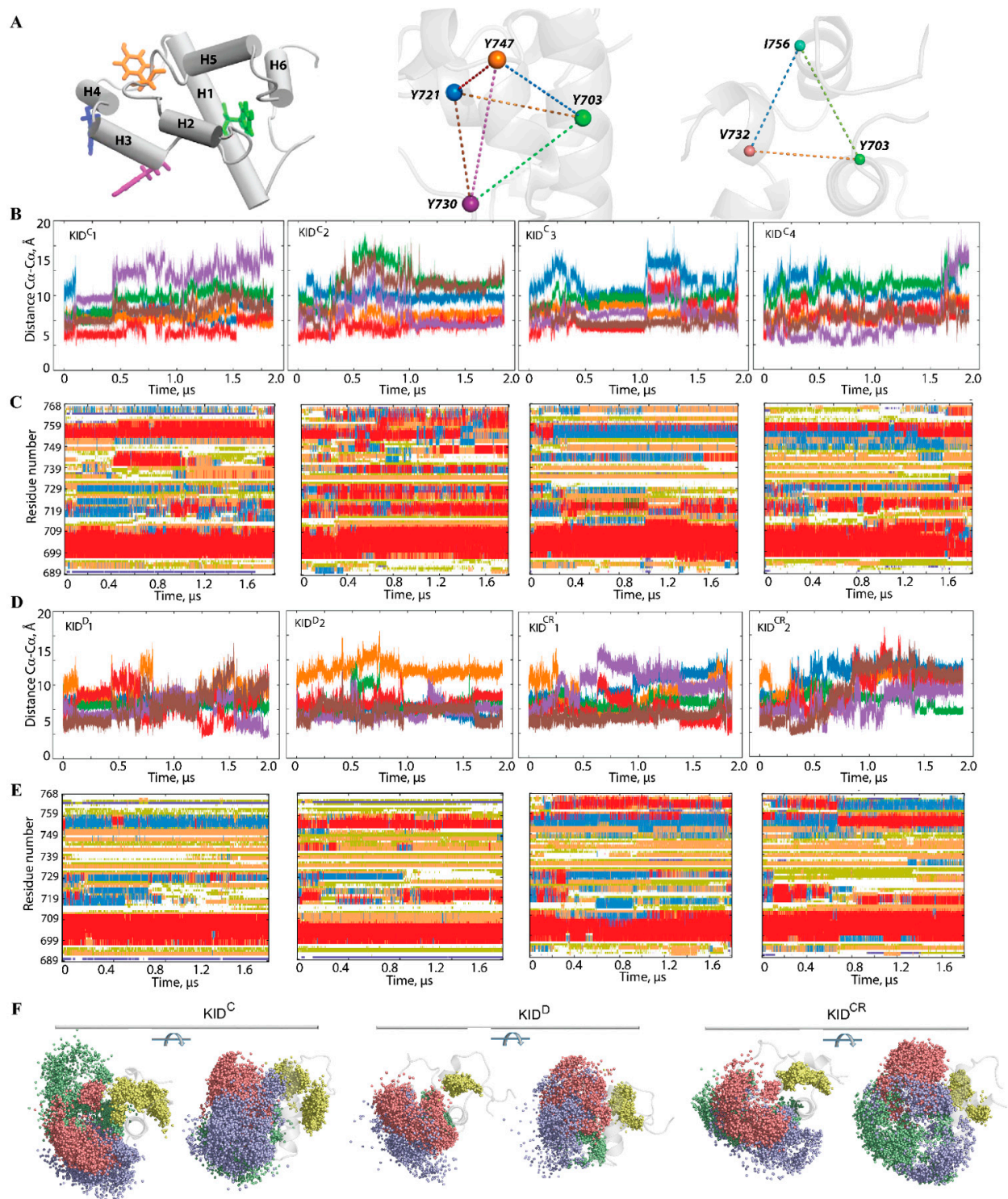### 2.5. Geometry of the Tyrosine Residues in KID

As KID contains four tyrosine residues, three of which are known to be phosphorylation sites (Y703, Y721, and Y730) and one (Y747) that has an unclear functional role [25], we focused on their structural features. These residues do not belong to the helices H1–H6 identified in KID, except Y703, which is positioned on a highly conserved αH1-helix. The other phosphotyrosine residues are located on transient fragments converting between α-, 3_{10}-helix and random coil. This observation suggests that the higher solvent accessibility of the phosphorylation sites afforded by the absence of secondary structure facilitates the post-transduction (translational) modifications required for recognition and recruitment of downstream molecules that are adaptors or signalling proteins.

We suggest that the geometry of the phosphotyrosine residues, key elements for substrate binding, may reflect (recover) the structural or conformational features of KID. Geometry was described by using a tetrahedron designed on the Cα-atoms of tyrosine residues regarded as nodes connected by edges (Figure 5). In the studied KID entities, the tetrahedron geometry varies greatly during the MD trajectories. Even if the edges of the tetrahedron are maintained over an extended period (30–50 ns), their lengths are further altered, either for all edges (synchronous transform) or only for certain edges (asynchronous transform).

Changes of the tetrahedron geometry during an MD trajectory are viewed either as instantaneous events or as stepwise processes. The long-time preservation of inter-tyrosine distances is apparently associated with the maintenance of the secondary structure in KID, while the synchronous/asynchronous change reflects the important transformation at the level of the helices. Consequently, the inter-tyrosine geometry is closely related to the folding–unfolding process in KID. Nevertheless, the other structural factors, such as relative orientation of the helices and flexibility of the coiled linkers, can contribute significantly to the high variability of tyrosine residue geometry. A coherent consequence of such variability of geometry of the tyrosine residues is the great dispersion of the hydroxyl groups (Figure 5F).

Furthermore, we focus on finding geometrically conserved elements of KID and their spatial relationships with incongruent structures. We suggest that residues Y703 (αH1-helix), V732 (H5 helix) and I756 (linker connecting H4 and H5 helices), with the smallest RMSF values, are a 'rigid subset' if the inter-residue distance is conserved in MD conformations. As expected, the inter-residue distances display essentially small variations compared to the distances between the tyrosine residues (Figure S7). Nevertheless, the almost invariant geometry of the most 'rigid' residues over the large periods of simulation time (≥ 1 μs) is followed either by an instantaneous change or by staggering alternations of their values. There is no obvious correlation between the geometry of a 'rigid subset' formed by Y703, V732 and I756, and the geometry of a tetrahedron designed on the tyrosine residues, Y703, Y721, Y730 and Y747; nonetheless, their changes without a doubt are connected.

Figure 5. The inter-residue geometry of tyrosines in the isolated unconstrained KID and its relationship with folding. (**A**) KID structure with helices shown as solid cylinders and tyrosine residues as sticks (left), tetrahedron delimited on the Cα-atoms of tyrosine residues (middle) and triangle designed on the most stable among the MD simulation residues (with the smallest RMSF values) (right). (**B,D**) Distances between each pair of tyrosine residues in each MD trajectory, coloured as the edges of the tetrahedron. (**C,E**) The time-related evolution of the secondary structures of each residue as assigned by DSSP with the type-coded secondary structure bar. (**D**) Distances between the most 'stable' (minimal RMSF values) residues Y703, V732 and I756 over each cMD trajectory, coloured as the edges of the triangle in (**A**). (**F**) The spatial distribution of hydroxyl groups presented by the oxygen atoms of the tyrosine residues in KID is shown in two orthogonal projections with the oxygen atoms of Y703, Y721, Y730 and Y747, respectively coloured in yellow, blue, red and green.

2.6. Free Energy Landscape as a Quantitative Measure of Folding and Disorder in KID

A promising strategy for the in-depth analysis of a protein conformational space is to consider the 'free energy landscape' along specifically chosen coordinates called 'reaction coordinates' or 'collective variables', which describe the conformation of a protein [26–28]. Such interpretation leads to quantitatively significant results that allow comparison between different states of a protein. The relative Gibbs free energy ΔG between two or more states is a measure of the probability of finding the system in those states. Such representations of protein sampling with use of reaction coordinates can be the quintessential model system for barrier crossing events in proteins [29]. These can be estimated from incomplete sampling of the states, as long as it is an unbiased sampling.

The statistical quantities—radius of gyration, RMSD, helical folding, contacts, surface exposed to solvent—usually used for the description of the conformational properties of proteins, were regarded as reaction coordinates (collective variables) for the evaluation of the relative free energy (ΔG) and reconstruction of its landscape (Figure 6).

First, the probability (P) of Rg, $P(Rg)$, and of RMSD, $P(\text{RMSD})$, were used as the reaction coordinates for the evaluation of the relative free energy (ΔG). The free energy landscape (FEL) as a function of RMSD and radius of gyration Rg (FEL^{Rg}_{RMSD}) for the concatenated replicas of KID is shown for each KID entity.

Each FEL^{Rg}_{RMSD} shows a rugged landscape indicating a high conformational heterogeneity, which is best represented in the cleaved KID (Figure 6A). The heterogeneity likely builds up the entropy barrier and adds complexity to the free energy map, and thus limits spontaneous state-to-state transition when using conventional advanced sampling methods. Nevertheless, the FEL^{Rg}_{RMSD} of each KID entity shows areas of minimum energy indicated by the red colour. The red areas represent more stability, while the reddened areas indicate transitions in the conformation of the protein followed by the thermodynamically more favourable state. The FEL^{Rg}_{RMSD} of different entities of KID have a unique global energy minimum (in dark red), which is clearly different from the other local minima showed by higher energy values (four minima in KID^C and two in KID^D and KID^{CR}). Regarding the FEL^{Rg}_{RMSD} ranges, it becomes the narrowest in KID^D and the largest in KID^C.

To characterise the KID conformations in terms of compactness, despite very small discrepancies in the values of Rg, the KID conformations were defined, based on mean values, as compact (11.2–11.6 Å), semi-compact (11.7–12.2 Å) and loose (12.3–12.8 Å). In KID^D, the lowest energy well is constituted of KID conformations having a semi-compact structure (Rg of 11.83(1) Å) (Figure 6B, Table S1). Two other wells are composed of loose conformations showing very close degrees of compaction (Rg of 12.2 and 12.3 Å).

In KID^{CR}, the narrow well of the lowest energy includes the KID conformations with a compact structure (Rg of 11.33(1) Å). Two other wells contain compact and semi-compact conformations (Rg of 11.2 and 11.8 Å, respectively). The low energy well on the large-ranged FEL^{Rg}_{RMSD} map of KID^C is composed of compact conformations (Rg of 11.80(1) Å); the other wells are composed of conformations having different compactness, with Rg varying from 11.4 Å (compact) to 12.8 Å (loose).

Second, the relative free energy ΔG of KID was evaluated while using as reaction coordinates the helical folding order parameter (H_{fp}), describing the fraction of helices (as calculated by DSSP), and the radius of gyration (Rg). The unimodal Gaussian distribution of H_{fp} does not separate the KID conformations on isolated clusters but describes the range of variation of the helical content and delimits the most populated region (Figure 6C). The H_{fp} and Rg values are apparently decorrelated. The free energy landscape, $FEL^{Rg}_{H_{fp}}$, reconstructed on these coordinates, shows a unique well for each KID entity, but their profile and position is differed between the studied KID (Figure 6D). This unique deep well, large in KID^C and narrow in KID^D and KID^{CR}, is composed of conformations that show a different order of helical folding that is greater in cleaved KID (KID^C and KID^{CR}) compared to the KID domain of KIT (KID^D).
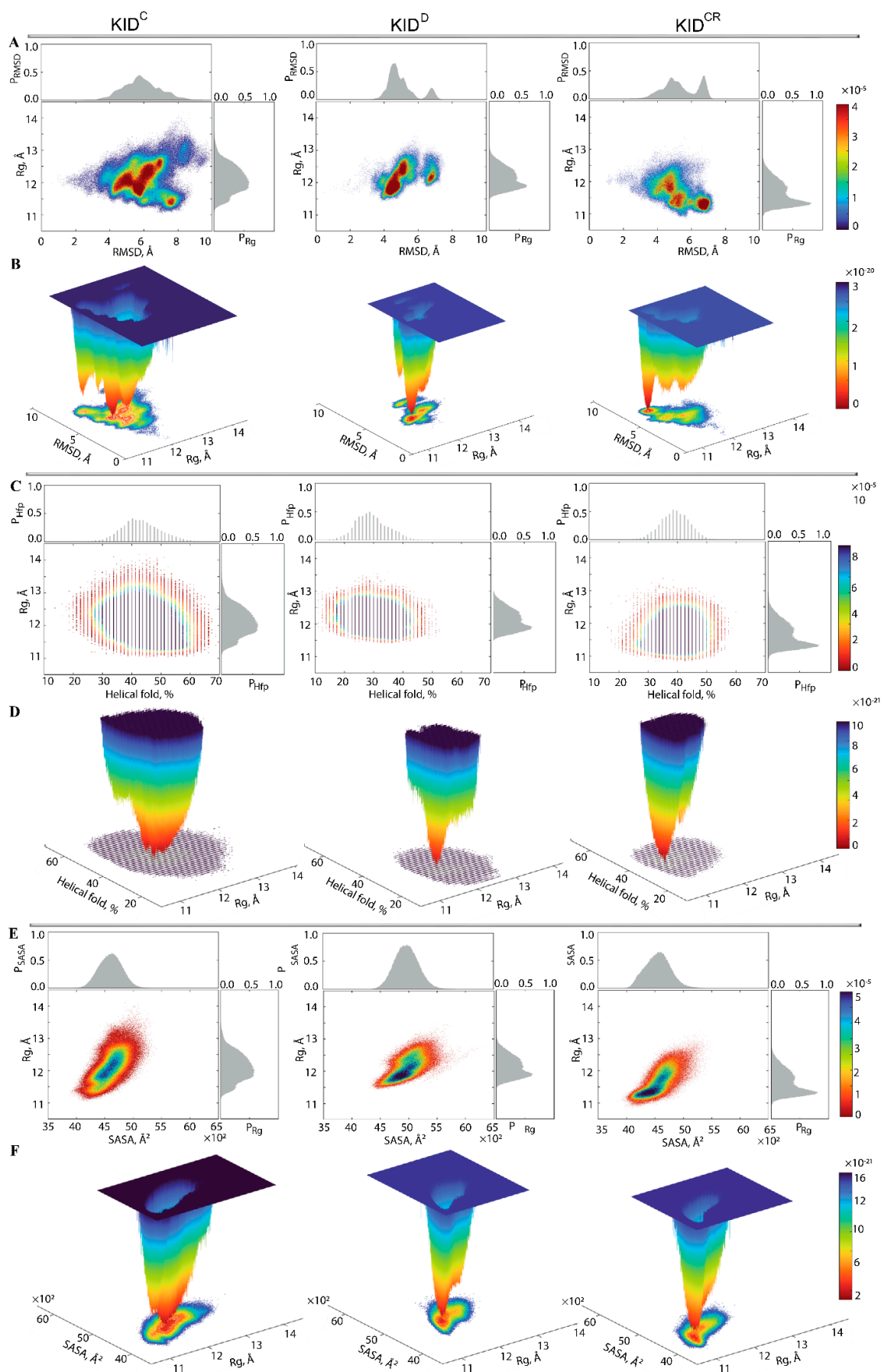
Figure 6. Free energy landscape (FEL) of KID as a function of the reaction coordinates. Each FEL generated on the MD

conformations of different KID entities (KID^C, KID^D and KID^{CR}) is displayed for the conformational ensemble sampled on the merged replicas. Calculations of the radii of gyration are performed on the Cα-atoms. The 2-dimensional representation of FEL of the KID conformational ensembles plotted as a function of Rg (in Å) versus (**A**) RMSD (in Å); (**C**) H_{fp}, which describes the fraction of helices (as calculated by the DSSP) (in %); (**E**) SASA, which describes the solvent-accessible surface area (as calculated with Cpptraj) (in Å^2). Probability distribution of each reaction coordinate is shown at the top and right, respectively. The red colour represents high occurrence, yellow and green low, and blue represents the lowest occurrences. The 3-dimensional representation of the relative Gibbs free energy as a function of Rg and (**B**) RMSD; (**D**) H_{fp}; (**F**) SASA. The blue colour represents the high energy state, green and yellow low and blue represents the lowest stable state. The free energy surface was plotted using Matlab.

Third, for evaluation of the relative free energy (ΔG) of KID, the solvent-accessible surface area (SASA) and the radius of gyration (Rg) were used as reaction coordinates. Statistically, similarly to the helical content, the SASA is also represented by the unimodal Gaussian distribution, which is highly symmetric in KID^C and KID^D, and slightly asymmetric in KID^{CR}. The SASA and Rg are highly correlated metrics, showing the dependence of the solvent-accessible surface from the KID size. The free energy landscape FEL^{Rg}_{SASA}, reconstructed on these reaction coordinates, shows deep and shallow wells which are adjacent in each KID entity. In KID^C and KID^D these wells are composed of semi-compact KID conformations which are more exposed to solvent in KID^D and rather buried in KID^C. In KID^{CR}, the deeper well is populated by the compact and buried conformations.

The free energy landscape of KID, represented as a function of the radius of gyration Rg and of an alternate metric (RMSD, helical folding parameter, SASA) as the second reaction coordinate, groups the KID conformations in a statistically found collection of multi-minima, or super-wells. On each FEL we can distinguish the most probable states of different KID entities and compare their functionally related properties. We can also understand how these statistical properties vary in different parts of the energy landscape.

As emerges from the observations mentioned above, the representative conformations of the deepest well (the lowest energy minimum) located on each FEL do not show close similarity in their folding (secondary structure), tertiary organisation (3D structure) or the spatial position of tyrosine residues for the same KID entity (Figure 7, Table S1). Nevertheless, in KID^D these randomly chosen conformations cognate well (RMSD of 2.0–3.1 Å), while in the cleaved KID, KID^C and KID^{CR}, they are largely different (Figure S8, Table S1).

The conformational sub-set of the deepest well on each FEL indicated the homogenous composition for each KID entity, as shown by similar first-principles metrics, RMSD and Rg, varying a little. It seems interesting to compare the similarity of the conformational subset from the deepest well of each KID entity with those obtained by ensemble-based clustering. The Rg mean values of the same KID entity computed on two conformational subsets are equivalent. This observation is valid for all the KID species studied. Comparing the most populated conformational subsets generated by ensemble-based clustering with those of the deepest wells, we suggest their agreement, although very approximate. Obviously, each deepest well on FEL^{Rg}_{RMSD} is formed by the most similar KID conformations of close size then the most populated cluster contained the more heterogeneous conformations.
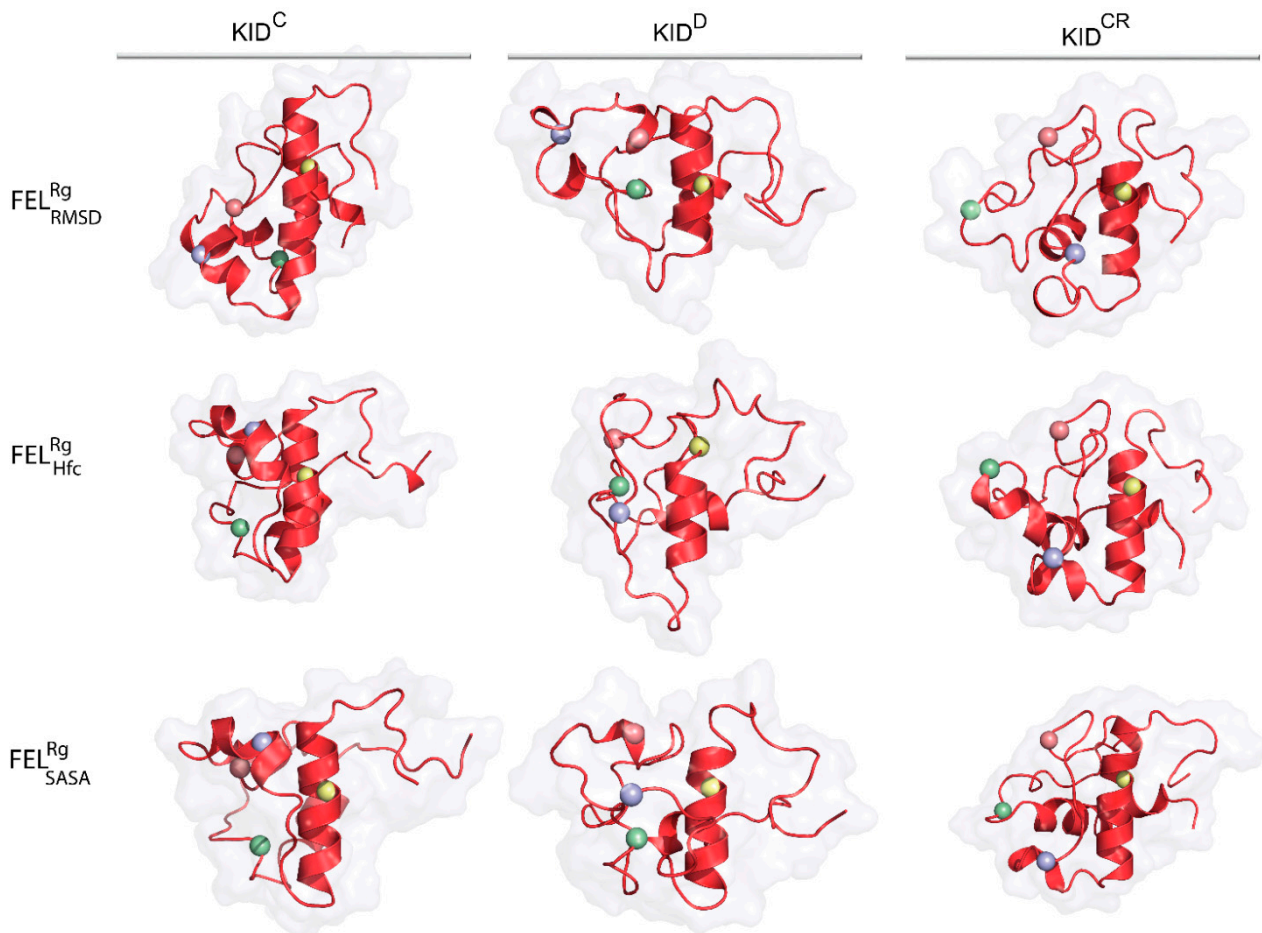
Figure 7. Representative conformations of the deepest well on the free energy landscape (FEL) of KID. KID displayed as red ribbons contoured with a surface-filled model. Position of the tyrosine residues (Cα-atoms) is shown as balls coloured in yellow, blue, red and green for Y703, Y721, Y730 and Y747, respectively.

3. Discussion

In the present study, we focused on the kinase insert domain of RTK KIT, the key recruitment region for a host of downstream signalling proteins—enzymes and adapter/scaffolding proteins [30]. Using the MD simulation-based study, we showed that the KIT kinase insertion domain is intrinsically disordered, and this property is disconnected from the KID context, as a cleaved polypeptide or as a domain of the receptor. Intrinsic disorder is a central feature in the function of numerous proteins, enabling them to serve as molecular switches and hubs in biological networks [31,32]. The intrinsically disordered regions (IDRs) of these proteins are increasingly recognised for their prevalence and their critical roles in regulatory intermolecular interactions. It has been estimated that IDRs in the human proteome contain ~132,000 binding motifs [33]. Disordered proteins are believed to account for a large fraction of all cellular proteins, playing roles in cell-cycle control, signal transduction, transcriptional and translational regulation, and large macromolecular complexes [34].

Principally, the RTK KIT contains three regions, JMR, KID and C-terminal, which retain the phosphotyrosine residues regulating signalling in multiple ways. The activation loop (A-loop) of KIT holds also two tyrosine residues, Y812 and Y823, the functions of which are still under discussion [35,36]; however, it was demonstrated that the mutation of Y823 causes aberrant downstream signalling and a possible binding of proteins by the phosphorylated Y823 was suggested [36]. Consequently, these four phosphotyrosine-containing

regions of KIT are the principal regulatory subunits, maintaining the spatiotemporal control of KIT signalling.

In KIT, these regions are mainly composed of polar and charged residues, reaching up to 58% in KID and C-terminal, 46% in JMR and 43% in A-loop, while the portion of hydrophobic residues is reduced to 34% in JMR and KID, and 36% in C-terminal, while in A-loop their amount reach up to 47%. High content of polar and charged residues and a low population of hydrophobic residues is an archetypal signature of the intrinsically disordered proteins [21,37]. The lower portion of charged and polar residues in JMR and A-loop compared to KID and C-term, is reflected in the enhanced stability of their 2D and 3D structure in each receptor's state, inactive or active, as shown in crystallographic structures [38,39]. At the same time, A-loop and JMR, compared to the other KIT regions, show the largest difference in their position and structure in the inactive and active states (RMSD of 17 and 10 Å for A-loop and 5 and 2 Å for JMR), suggesting their intrinsic disorder. While KID is longer in sequence and has a higher fraction of disordered residues compared to the other phosphotyrosine-containing KIT regions, its intrinsic disorder is more abundant.

The intrinsic disorder of KID makes it difficult to study by experimental techniques [40], and the 3D numerical model is the unique description of KID from KIT [11]. By taking these 3D data as the initial structure for the study of its evolution over time, we obtained an accurate and detailed atomistic physical model which presents the intrinsically disordered kinase insertion domain of KIT as a set of heterogeneous conformations. The ensembles of conformations generated for each studied KID entity were characterised by using statistical descriptors which are tightly related with either the geometrical (structural) features or the physical properties.

Our results show that accurate physical models of flexible biomolecular systems, such as proteins consisting of several intrinsically disordered regions, are well founded and can serve to pave the way to establishing the relationship between conformational flexibility and biological function.

The KID folding displays a collection of highly variable helices altered in length and type (α- or 3_{10}-helices). Except for H1, a unique helix constantly folded as an α-helix and varying only in its length, the other KID helices are transient, converting between the α- and 3_{10}-helices or a helix and non-regular secondary structures—turns, bends, and coils. Apparently, the helical fold is a proper feature of KID, disconnected from KID status (as a cleaved species or a domain fused to KIT) and of a context of MD simulations (unconstrained or restrained distance F689-D768). Nevertheless, a quantity of such folding depends partially not only on the restriction applied to KID extremities, but also on the kind of such restriction. First, a greater degree of the 2D folding was detected in the cleaved KID (KID^C and KID^{CR}), evidenced by the arrangement of transient helices H4 and H6, not observing in KID fused to KIT (KID^D). Moreover, comparing 2D structures of the cleaved KID, we observed that H2 and H3 helices are more extended in KID simulated with restrained N- and C-ends (KID^{CR}).

The conserved α-helical architecture of the H1 helix, reported for KID embedded to KIT [11], is also observed in each studied KID entity, independently from its context—KID fused to KIT or cleaved polypeptide. Such structural consistency, together with conservation of its spatial position, suggests that this helix, immediately adjacent to the kinase domain, is the inner ordered motif of KID, which is critical for KIT function. Indeed, despite the changes in the helix length along the same trajectory or in different trajectories of distinct KID entities, H1, like a drop of glue, which attaches the KID structural fragments around them by manifold contacts and stabilises a 'globule-like' shape of the intrinsically disordered KID. The 'organising role' in stabilising the KID structure was previously attributed to tyrosine Y747 located on the helix H4 [11]. We suggest that the Y747 and H1-helix functions are complementary and can be mutually dependent.

The reduced portion of folded secondary structure in KID fused to KIT (KID^D) suggests that its folding is controlled by the kinase domain. We postulate that the KID folding

is principally sequence dependent, but partly allosterically regulated by the kinase domain of KIT. Allostery within or mediated by intrinsically disordered proteins ensures robust and efficient signal integration through mechanisms that would be extremely unfavourable or even impossible for ordered globular protein upon its interaction with the partners [41,42].

The other inherent structural feature of KID is related to flexible linkers connecting these helices, which favour many relative orientations of helices, leading to a vast set of highly heterogeneous conformations. Apparently, the conformational flexibility may be more important than the specific secondary structures before binding. The length of KID linkers is variable and strongly depends on the order of helical folding. Our study revealed that despite great flexibility, once in either a cleaved state or fused to KIT, KID acquires a compact globule-like shape, with a helical fold fraction ranging from the poor (in KID^D) to rich (KID^C). A helical structure is, in general, not stable by itself—short helices taken out from stable globular proteins are found to be unstable when isolated [43]. Additional stabilising interactions of the helical structure must, therefore, be present in globular proteins.

In KID, the bulky hydrophobic residues of the weakly fluctuating regions localised on ordered structures (H1 and H5) or random coil (linker between H3 and H4/H5), interact strongly between themselves, forming a ternary hydrophobic core which stabilises the entire tertiary structure of the helically folded KID. The stabilisation of the N- and C-termini of KID is achieved by switchable H-bonds formed by terminal residues in various combinations in different KID entities.

The intrinsically disordered KID contains three known phosphorylation sites (residues Y703, Y721 and Y730) that control KIT signalling. Tyrosine Y703 is located on the αH1-helix, the most conserved structure compared to the rest of KID, although its length varies greatly in each KID entity. The other phosphorylation sites are located on fully transient structures. These observations raised the possibility that, when targeting signalling proteins, KID could display a rich landscape of overlooked folding/dynamic properties. We suggest that (i) transient helices are structures preconfigured to specifically localise signalling proteins, which are selective for alternative phosphotyrosine sites of KID, and may facilitate phosphorylation-mediated regulation of signalling cascade, and (ii) KID has evolved an unusual structural flexibility to stabilise long-lived folding intermediates and, thus, maximise their 'recognisability' to signalling proteins. The specificity of intermolecular interactions of KID with signalling proteins is apparently determined by sequence- and structure-based selectivity, which are the two determining factors in 'molecular recognition'. Furthermore, the transient helices provide inherent stability near phosphorylation sites in KID, consistent with the finding of short and transient helical structures near the phosphorylation sites of the (unphosphorylated) disordered region of the Src homology 2 protein domains (SH2) of adaptor or signalling molecules—Grb2, PI3K, PLCγ [5].

Several questions/issues naturally arise regarding fundamental aspects of the folding/unfolding process in KID, such as the following: (i) explicitly characterising the free energy landscape of folding, which, as we found, is coupled to the KID status and naturally to its binding properties; (ii) to choose topographic characteristics as a criterion for referencing a cleaved KID as the appropriate species for the study of post-transduction process. Furthermore, it is of significant interest to explore to what extent the funnel landscape perspective also applies to intrinsically disordered proteins.

The first principles describing the protein properties—the size (the radius of gyration, Rg), the RMSD (the measure of the average distance between the conformations), the amount of the organised secondary structure (the helical folding order parameter, Hel_{fp}) and the solvent accessible surface area (SASA)—were used as reaction coordinates (collective variables) for the evaluation of the relative free energy (ΔG) of KID. Similar measurements have been used to compute interesting protein physics [44–46]. Furthermore, such first-principles-based descriptions of the free energy landscape of a disordered protein could be a part of machine learning algorithms that attempt to find energy functions that will finely represent the protein landscape.

The native structure of a protein, as stated in free energy theory [47], shows the lowest free energy in the large conformational space. In this case, the free energy landscape of protein folding is funnel shaped and oriented towards a single well (basin) corresponding to a native structure. The 'trapping' conformations observed during the folding process show a shallower well depth than the overall bias towards the native structure, which ensures that the native structure is both thermodynamically favourable and kinetically accessible. This is valid in the case of a well-ordered protein, whereas the IDP we suggest has several 'quasi-native' states. Despite some conventionality in using RMSD as a reaction coordinate, the free energy landscape defined on Rg and RMSD showed a series of well-resolved wells, which can be interpreted as 'quasi-native' states of intrinsically disordered KID. Landscapes defined on the basis of structurally or physically related measures (SASA, Rg or order of helicity), used as reaction coordinates, displayed a single deep and enlarged well.

We show that the cleaved unconstrained KID explores multiple conformational sub-states, which are represented by conformations ranging from compact to lose, regardless of their helical content, and showed that the solvent-accessible surface area is highly dependent on the radius of gyration. KID with ends naturally restricted in KID^D or mimicking a native restriction in KID^{CR} exhibit a detectable reduction in helical structure and reduced conformational variability (diminished number of substates). Most of the conformations forming the deepest well on the FELs of KID^{CR} are characterised by high compactness and reduced solvent accessible surface area, while in KID^D, this well is completed by semi-compact conformations that have a larger solvent-accessible surface area than in KID^{CR} and comparable with those in KID^C. In general, KID has been found to occupy numerous conformational substates in the weakly funnelled free energy landscape, which distinguishes it from the highly directed, time-dependent free energy landscape of regularly folded globular proteins [48–50].

The free energy landscapes constructed from first principles describe KID conformations through a collection of minima (so-called basins) and provide a direct 'polymetric' evaluation of conformational ensemble each KID species and a comparison between the conformational ensembles of different KID entities. This description is much more accurate than grouping conformations using ensemble-based clustering and allows direct comparison between proteins based on structural metrics or physically related parameters.

By comparing the respective species, we found that the profiles of the free energy landscapes of KID^D and KID^{CR} are approximately identical, with only one deep well, while in the unconstrained cleaved KID the free energy landscape is highly disrupted, which leads to slightly different wells with clearly pronounced minima. This indicates that the conformational ensemble of the cleaved KID simulated using restricted N- and C-termini (KID^{CR}), better reproduces the KID from the KIT.

Therefore, for studying post-transduction events of KIT, KID^{CR} is the most suitable entity. We suggest that the N- and C-termini of the cleaved KID can be stabilised by a linker, which is a chemical or polypeptide fragment 10 Å in length. Such a cyclic polypeptide would be a *bona fide* generic molecule mimicking the native KID for both computational and empirical studies of KID phosphorylation and recruitment of a multitude of signal transducers to its docking sites. Nevertheless, it seems to us that the use of the cleaved KID cannot be universal. Even for the study of individual downstream signalling pathways mediated, for example, by KIT possessing 'oncogenic driver' mutations in the tyrosine kinase domain, a full-length cytoplasmic domain would be required.

4. Materials and Methods

4.1. 3D Models

The 3D model of the full-length cytoplasmic domain of KIT was taken from [11]. The coordinates of kinase insert domain (KID, sequence F689–D768 aas) were extracted and used as a model of the cleaved KID.

4.2. Molecular Dynamics Simulation

4.2.1. Preparation of the Systems

For MD simulation, models of KIT and KID (inactive unphosphorylated state) were prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBER) [51] using the *ff99SB* all-atom force field parameter set: (i) hydrogen atoms were added, (ii) covalent bond orders were assigned, (iii) protonation states of amino acids were assigned based on their solution for pK values at neutral pH, (iv) histidine residues were considered neutral and protonated on ε-nitrogen atoms, and (v) Na^+ counter-ions were added to neutralise the protein charge.

Each protein was solvated with explicit TIP3P water molecules in a periodic rectangular box with at least 12 Å distance between the proteins and the boundary of the water box. The total number of atoms in the systems (protein, water molecules and counter ion) was 69,089 and 19,537 for the KIT and KID, respectively.

4.2.2. Setup of the Systems

The setup of the systems was performed with the Simulated Annealing with NMR-Derived Energy Restraints (SANDER) module [52] of AMBER. First, each system was minimised successively using the steepest descent and conjugate gradient algorithms as follows: (i) 10,000 minimisation steps where the water molecules have fixed protein atoms, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein side chains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. After relaxation, each system was gradually heated from 10 to 310 K at constant volume using the Berendsen thermostat [53] while restraining the solute Cα atoms by 10 kcal/mol/Å^2. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for a further 100 ps at constant pressure (NPT) maintained by the Monte Carlo method [54]. Final system equilibration was achieved by a 100 ps NPT run to assure that the water box of the simulated system had reached the appropriate density. The electrostatic interactions were calculated using the PME (particle mesh Ewald summation) [55] with a cut-off of 10.0 Å for long-range interactions. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution.

4.2.3. Production of the MD Trajectories

All trajectories were produced using the AMBER ff99SB force field with the PMEMD module of AMBER 16 and AMBER 18 [51] (GPU-accelerated versions) running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2× CPU Intel Xeon E5-2680 and Nvidia GTX 780ti) and the supercomputer JEAN ZAY at IDRIS.

The multiple extended trajectories were generated for each equilibrated system: two 2-μs trajectories for KIT with KID, four 1.8-μs replicas for cleaved KID (KID^C) and two 1.8-μs replicas for cleaved KID with the restrained distance (10 Å) between the Cα-atoms of terminal residues, F689 and D768 (KID^{CR}).

A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog method. Coordinate files were recorded every 1 ps. Neighbour searching was performed by the Verlet algorithm [56]. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step. The van der Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution. Coordinates were recorded every 1 ps.

4.2.4. Data Analysis

Unless otherwise stated, all recorded MD trajectories, individual and merged, were analysed (RMSFs, RMSDs, DSSP, clustering) with the standard routines of the CPPTRAJ 4.15.0 program [57] of AMBER 20 Suite. All analysis was performed on the MD conformations (every 10 ps) by considering either all simulations or the production part of the simulation, which was generated after the removal of non-well-equilibrated conformations

(0–70 ns), as was shown by the RMSDs, or on residues with a fluctuation of less than 4 Å, as shown by the RMSFs, and after least-square fitting [58,59] of the MD conformations for a region of interest, thus removing rigid-body motion from the analysis.

(1) The RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference.

(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method [60]. The secondary structure types were assigned for residues based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 and 20 ps for the individual and concatenated trajectories, respectively.

(3) Clustering analysis was performed on the productive simulation time of each MD trajectory using an ensemble-based approach [24]. The analysis was performed every 100 ps. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cut-off r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD con-formations are grouped into n reference clusters based on their RMSDs from each reference structure. The cut-off was varied from 3 to 5 Å.

(4) H-bonds between heavy atoms (*N*, *O*, and *S*) as potential donors/acceptors were calculated with the following geometric criteria: donor/acceptor distance cut-off was set to 3.6 Å, and the bond angle cut-off was set to 120°. Hydrophobic contacts were considered for all hydrophobic residues with side chains within a 4 Å of each other.

(5) Contact search was performed with Cpptraj. Contact present if Cα–Cα distance <10 Å. Map normalised over the number of frames such as the value represents the contact frequency in the considered trajectory.

(6) The radius of gyration (Rg) was calculated from the atomic coordinates using the formula (Equation (1)) from [61]:

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}} \tag{1}$$

where m_i is the mass of the atom i, r_i is the distance of atom i from the center of mass of the protein.

(7) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates (Equation (2)) [62]:

$$\Delta G_{(R1,R2)} = -k_B T \ln \frac{P_{(R1,\ R2)}}{P_{max}} \tag{2}$$

where k_B represents the Boltzmann constant, T is the temperature. $P_{(R1,\ R2)}$ denotes the probability of states along the two reaction coordinates, which is calculated using a k-nearest neighbour scheme and P_{max} denotes the maximum probability. The 3-dimensional representations of the free energy surface were plotted using Matlab (© 1994-2021 The MathWorks, Inc., Natick, MA, USA).

4.2.5. Advanced Data Analysis

Secondary structure for each KID residue of each replica (DSSP) was classified by an 8-letter code [63] and used for calculation of an estimated transition probability matrix from one folding state to another. A suitable distance (Fisher-Rao distance) between these families was calculated for all the pairs of replicas. Multi-Dimensional Scaling (MDS) was performed to get an 'as isometric as possible' embedding of the data in 2D (i.e., a

representation by placing points on a plane while preserving the calculated inter-distances as well as possible).

4.2.6. Visualisation and Figure Preparation

Visual inspection of the conformations and figure preparation were performed with PyMOL (https://pymol.org/2/, accessed on 14 September 2020). The VMD 1.9.3 program (accessed on 21-12-2020) [64] was used to prepare the protein MD animations. To visualise the motions along the principal components, the Normal Mode Wizard (NMWiz) plugin [65], which is distributed with the VMD program, was used.

References

1. Lemmon, M.A.; Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **2010**, *141*, 1117–1134. [CrossRef]
2. Volinsky, N.; Kholodenko, B.N. Complexity of receptor tyrosine kinase signal processing. *Cold Spring Harb. Perspect. Biol.* **2013**, *5*, a009043. [CrossRef]
3. Changeux, J.-P.; Christopoulos, A. Allosteric Modulation as a Unifying Mechanism for Receptor Function and Regulation. *Cell* **2016**, *166*, 1084–1102. [CrossRef] [PubMed]
4. Schlessinger, J. Cell signaling by receptor tyrosine kinases. *Cell* **2000**, *103*, 211–225. [CrossRef]
5. Rönnstrand, L. Signal transduction via the stem cell factor receptor/c-Kit. *Cell. Mol. Life Sci.* **2004**, *61*, 2535–2548. [CrossRef] [PubMed]
6. Pham, D.D.M.; Guhan, S.; Tsao, H. KIT and Melanoma: Biological Insights and Clinical Implications. *Yonsei Med. J.* **2020**, *61*, 562–571. [CrossRef] [PubMed]
7. Gilreath, J.A.; Tchertanov, L.; Deininger, M.W. Novel approaches to treating advanced systemic mastocytosis. *Clin. Pharmacol.* **2019**, *11*, 77–92. [CrossRef]
8. Kitamura, Y.; Hirotab, S. Kit as a human oncogenic tyrosine kinase. *Cell. Mol. Life Sci.* **2004**, *61*, 2924–2931. [CrossRef]
9. Ghosh, S.; Marrocco, I.; Yarden, Y. Roles for receptor tyrosine kinases in tumor progression and implications for cancer treatment. *Adv. Cancer Res.* **2020**, *147*, 1–57. [CrossRef]
10. Oved, S.; Yarden, Y. Signal transduction: Molecular ticket to enter cells. *Nature* **2002**, *416*, 133–136. [CrossRef]
11. Inizan, F.; Hanna, M.; Stolyarchuk, M.; Chauvot de Beauchêne, I.; Tchertanov, L. The First 3D Model of the Full-Length KIT Cytoplasmic Domain Reveals a New Look for an Old Receptor. *Sci. Rep.* **2020**, *10*, 5401. [CrossRef] [PubMed]

12. Schramm, A.; Bignon, C.; Brocca, S.; Grandori, R.; Santambrogio, C.; Longhi, S. An arsenal of methods for the experimental characterization of intrinsically disordered proteins—How to choose and combine them? *Arch. Biochem. Biophys.* **2019**, *676*, 108055. [CrossRef] [PubMed]

13. Kasahara, K.; Terazawa, H.; Takahashi, T.; Higo, J. Studies on Molecular Dynamics of Intrinsically Disordered Proteins and Their Fuzzy Complexes: A Mini-Review. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 712–720. [CrossRef]

14. Amit, I.; Wides, R.; Yarden, Y. Evolvable signaling networks of receptor tyrosine kinases: Relevance of robustness to malignancy and to cancer therapy. *Mol. Syst. Biol.* **2007**, *3*, 151. [CrossRef]

15. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* **2017**, *7*, 818–831. [CrossRef]

16. Bu, Z.; Callaway, D.J.E. Proteins move! Protein dynamics and long-range allostery in cell signaling. *Adv. Protein Chem. Struct. Biol.* **2011**, *83*, 163–221. [CrossRef]

17. Changeux, J.P.; Edelstein, S.J. Allosteric mechanisms of signal transduction. *Science* **2005**, *308*, 1424–1428. [CrossRef]

18. Tsai, C.J.; Del Sol, A.; Nussinov, R. Protein allostery, signal transmission and dynamics: A classification scheme of allosteric mechanisms. *Mol. Biosyst.* **2009**, *5*, 207–216. [CrossRef]

19. Jadwin, J.A.; Ogiue-Ikeda, M.; Machida, K. The application of modular protein domains in proteomics. *FEBS Lett.* **2012**, *586*, 2586–2596. [CrossRef]

20. Paul, M.D.; Hristova, K. The RTK Interactome: Overview and Perspective on RTK Heterointeractions. *Chem. Rev.* **2019**, *119*, 5881–5921. [CrossRef]

21. Uversky, V.N. Introduction to Intrinsically Disordered Proteins (IDPs). *Chem. Rev.* **2014**, *114*, 6557–6560. [CrossRef]

22. Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V.N. Introducing protein intrinsic disorder. *Chem. Rev.* **2014**, *114*, 6561–6588. [CrossRef]

23. Oldfield, C.J.; Dunker, A.K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **2014**, *83*, 553–584. [CrossRef]

24. Lyman, E.; Zuckerman, D.M. Ensemble-based convergence analysis of biomolecular trajectories. *Biophys. J.* **2006**, *91*, 164–172. [CrossRef]

25. Mabe, S.; Nagamune, T.; Kawahara, M. Detecting protein–protein interactions based on kinase-mediated growth induction of mammalian cells. *Sci. Rep.* **2014**, *4*, 6127. [CrossRef]

26. Banavali, N.K.; Roux, B. Free Energy Landscape of A-DNA to B-DNA Conversion in Aqueous Solution. *J. Am. Chem. Soc.* **2005**, *127*, 6866–6876. [CrossRef]

27. Pietrucci, F. Strategies for the exploration of free energy landscapes: Unity in diversity and challenges ahead. *Rev. Phys.* **2017**, *2*, 32–45. [CrossRef]

28. Hénin, J.; Fiorin, G.; Chipot, C.; Klein, M.L. Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *J. Chem. Theory Comput.* **2010**, *6*, 35–47. [CrossRef] [PubMed]

29. Appadurai, R.; Nagesh, J.; Srivastava, A. High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. *Nat. Commun.* **2021**, *12*, 958. [CrossRef]

30. Hubbard, S.R.; Miller, W.T. Receptor tyrosine kinases: Mechanisms of activation and signaling. *Curr. Opin. Cell. Biol.* **2007**, *19*, 117–123. [CrossRef]

31. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; et al. Intrinsically disordered protein. *J. Mol. Graph. Model.* **2001**, *19*, 26–59. [CrossRef]

32. Wright, P.E.; Dyson, H.J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **2015**, *16*, 18–29. [CrossRef]

33. Wong, E.T.C.; So, V.; Guron, M.; Kuechler, E.R.; Malhis, N.; Bui, J.M.; Gsponer, J. Protein-Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations. *Biomolecules* **2020**, *10*, 1097. [CrossRef] [PubMed]

34. Dosztányi, Z.; Chen, J.; Dunker, A.K.; Simon, I.; Tompa, P. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* **2006**, *5*, 2985–2995. [CrossRef]

35. DiNitto, J.P.; Deshmukh, G.D.; Zhang, Y.; Jacques, S.L.; Coli, R.; Worrall, J.W.; Diehl, W.; English, J.M.; Wu, J.C. Function of activation loop tyrosine phosphorylation in the mechanism of c-Kit auto-activation and its implication in sunitinib resistance. *J. Biochem.* **2010**, *147*, 601–609. [CrossRef]

36. Agarwal, S.; Kazi, J.; Mohlin, S.; Påhlman, S.; Rönnstrand, L. The activation loop tyrosine 823 is essential for the transforming capacity of the c-Kit oncogenic mutant D816V. *Oncogene* **2015**, *34*, 4581–4590. [CrossRef]

37. Uversky, V.N. Intrinsically disordered proteins from A to Z. *Int. J. Biochem. Cell Biol.* **2011**, *43*, 1090–1103. [CrossRef]

38. Mol, C.D.; Dougan, D.R.; Schneider, T.; Skene, R.; Kraus, M.L.; Scheibe, D.N.; Snell, G.P.; Zou, H.; Sang, B.-C.; Wilson, K.P. Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem.* **2004**, *279*, 31655–31663. [CrossRef]

39. Mol, C.D.; Lim, K.B.; Sridhar, V.; Zou, H.; Chien, E.Y.T.; Sang, B.-C.; Nowakowski, J.; Kassel, D.B.; Cronin, C.N.; McRee, D.E. Structure of a c-kit product complex reveals the basis for kinase transactivation. *J. Biol. Chem.* **2003**, *278*, 31461–31464. [CrossRef] [PubMed]

40. Necci, M.; Piovesan, D.; Predictors, C.; Curators, D.; Tosatto, S.C. Critical Assessment of Protein Intrinsic Disorder Prediction. *bioRxiv* **2020**, *2020*, 245852. [CrossRef]

41. Berlow, R.B.; Dyson, H.J.; Wright, P.E. Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *J. Mol. Biol.* **2018**, *430*, 2309–2320. [CrossRef]

42. Hilser, V.J.; Thompson, E.B. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8311–8315. [CrossRef] [PubMed]

43. Dill, K.A. Dominant forces in protein folding. *Biochemistry* **1990**, *29*, 7133–7155. [CrossRef]

44. Onuchic, J.N.; Wolynes, P.G.; Luthey-Schulten, Z.; Socci, N.D. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 3626–3630. [CrossRef] [PubMed]

45. Onuchic, J.N.; Socci, N.D.; Luthey-Schulten, Z.; Wolynes, P.G. Protein folding funnels: The nature of the transition state ensemble. *Fold. Des.* **1996**, *1*, 441–450. [CrossRef]

46. Koretke, K.K.; Russell, R.B.; Lupas, A.N. Fold recognition without folds. *Protein Sci. Publ. Protein Soc.* **2002**, *11*, 1575–1579. [CrossRef] [PubMed]

47. Onuchic, J.N.; Wolynes, P.G. Theory of protein folding. *Curr. Opin. Struct. Biol.* **2004**, *14*, 70–75. [CrossRef]

48. Weinkam, P.; Zimmermann, J.; Romesberg, F.; Wolynes, P.G. The folding energy landscape and free energy excitations of cytochrome c. *Acc. Chem. Res.* **2010**, *43*, 652–660. [CrossRef]

49. Wolynes, P.G. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie* **2015**, *119*, 218–230. [CrossRef]

50. Chu, W.-T.; Wang, J. Quantifying the Intrinsic Conformation Energy Landscape Topography of Proteins with Large-Scale Open–Closed Transition. *ACS Sci.* **2018**, *4*, 1015–1022. [CrossRef]

51. Case, D.A.; Cheatham, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K.M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R.J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668–1688. [CrossRef]

52. Salomon-Ferrer, R.; Case, D.A.; Walker, R.C. An overview of the Amber biomolecular simulation package. *WIREs Comput. Mol. Sci.* **2013**, *3*, 198–210. [CrossRef]

53. Peters, E.J.F.; Goga, N.; Berendsen, H.J.C. Stochastic Dynamics with Correct Sampling for Constrained Systems. *J. Chem. Theory Comput.* **2014**, *10*, 4208–4220. [CrossRef] [PubMed]

54. Duane, S.; Kennedy, A.; Pendleton, B.; Roweth, D. Hybrid Monte Carlo. *Phys. Lett. B* **1987**, *195*, 216–222. [CrossRef]

55. Sagui, C.; Pedersen, L.G.; Darden, T.A. Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in biomolecular simulations. *J. Chem. Phys.* **2004**, *120*, 73–87. [CrossRef] [PubMed]

56. Hairer, E.; Lubich, C.; Wanner, G. Geometric numerical integration illustrated by the Störmer–Verlet method. *Acta Numer.* **2003**, *12*, 399–450. [CrossRef]

57. Roe, D.R.; Cheatham, I.T.E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095. [CrossRef] [PubMed]

58. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **1976**, *32*, 922–923. [CrossRef]

59. Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. Sect. A* **1978**, *34*, 827–828. [CrossRef]

60. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637. [CrossRef]

61. Lobanov, M.I.; Bogatyreva, N.; Galzitskaia, O.V. Radius of gyration is indicator of compactness of protein structure. *Mol. Biol.* **2008**, *42*, 701–706. [CrossRef]

62. Gapsys, V.; Michielssens, S.; Peters, J.H.; De Groot, B.L.; Leonov, H. Calculation of binding free energies. *Methods Mol. Biol.* **2015**, *1215*, 173–209. [CrossRef] [PubMed]

63. Konagurthu, A.S.; Lesk, A.; Allison, L. Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics* **2012**, *28*, i97–i105. [CrossRef] [PubMed]

64. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual molecular dynamics. *J. Mol. Graph.* **1996**, *14*, 33–38. [CrossRef]

65. Bakan, A.; Meireles, L.M.; Bahar, I. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **2011**, *27*, 1575–1577. [CrossRef]