Generalizable Protein Dynamics in Serine-Threonine Kinases: Physics is the key

Soumendranath Bhakat[1*], Shray Vats[1], Andreas Mardt[2], Alexei Degterev[3]

[1]AlloTec Bio, St. Louis, Missouri 63116, United States
[2]Independent Researcher, 12053 Berlin, Germany
[3]Department of Developmental, Molecular and Chemical Biology, School of Graduate Biomedical Sciences, Tufts University School of Medicine, Boston, Massachusetts, United States
*Email: sbhakat@allotec.bio /bhakatsoumendranath@gmail.com

Abstract

Serine-Threonine Kinases (STKs) play crucial roles in signaling pathways across oncology, inflammation, and neurodegenerative diseases. Historically, their conformational states have been defined by the DFG motif: DFG-in and DFG-out. However, this binary paradigm overlooks the broader conformational heterogeneity of apo STKs, which encompasses multiple metastable states within an expanded DFG-Phe ensemble. We introduced a novel nomenclature that integrates the dynamics of the DFG-Phe, activation loop, and αC-helix, highlighting how these regions respond to various ensemble perturbations (e.g., mutations, ligand binding, and protein–protein interactions). A major bottleneck in studying kinases lies in sampling their wide-ranging conformations because static snapshots often remain trapped in specific free energy minima, limiting traditional molecular dynamics simulations from exploring multiple functionally relevant states starting from a single structure. To accelerate conformational sampling, we introduce a computational framework that integrates AlphaFold, machine learning, physics-based simulations, and Markov state modeling. Rather than focusing on single-structure snapshots or structural hypotheses generated by protein structure prediction models, our framework captures shifts in conformational populations under varying perturbations, shedding light on both the thermodynamics and kinetics of the transitions. A key innovation lies in our machine learning algorithms, which capture slowly varying structural features to uncover hidden states and generate latent representations of conformational motions across different kinase domains. The physics-refined structural ensemble sampled from the latent layers is then used to launch new simulations that more comprehensively explore the full conformational landscape than traditional molecular simulation approaches. By capturing how these conformational shifts influence downstream protein–protein interactions, conformational allostery, and cryptic pocket formation, our accelerated simulation framework provides deeper insights into STK mechanisms and enables the development of new therapeutic modalities. Finally, by uniting our comprehensive conformational definition with an AI-accelerated molecular simulation strategy, we capture generalized conformational dynamics in kinases and how their heterogeneity is influenced by ensemble-level perturbations. This framework can be extended to investigate broader protein families—such as G protein-coupled receptors (GPCRs) and tumor necrosis factors (TNFs)—where functional outcomes are dictated by conformational heterogeneity.

Introduction

Serine/threonine kinases (STKs) are a large protein family with over 500 members in humans, responsible for phosphorylating serine and threonine residues on target proteins[1]. STKs participate in kinase-mediated protein–protein interactions that regulate downstream signaling pathways involved in oncology, inflammatory, and neurodegenerative diseases. The activity of STKs is heavily modulated by their conformational dynamics, which are influenced by perturbations such as mutations, small molecule binding, and protein–protein interactions[2,3].

Like most kinases, STKs contain a conserved Asp-Phe-Gly (DFG) motif essential for their function. Historically, kinase conformational dynamics have been described by the DFG-in and DFG-out states[4,5], corresponding to active and inactive conformations, respectively. In the active (DFG-in) conformation, the DFG-Phe residue is packed into a hydrophobic pocket, while the DFG-Asp faces outward. Conversely, in the inactive (DFG-out) state, the DFG-Asp flips, and the DFG-Phe relocates out of the hydrophobic pocket. For STKs, the transition from DFG-in to DFG-out requires overcoming a significantly high free energy barrier and is often induced by the binding of type II kinase inhibitors[6,7].

However, this binary classification overlooks the conformational heterogeneity of apo STKs, which predominantly exist in a dynamic equilibrium among multiple metastable states within a global DFG-in ensemble. The current nomenclature paradigm is inadequate to capture this conformational diversity. Understanding the populations and timescales associated with the conformational heterogeneity is essential for elucidating the effects of perturbations on STK dynamics, providing molecular-level insights into disease mechanisms, and facilitating the development of selective therapeutics targeting these kinases. Here, we propose a new nomenclature to define the conformational heterogeneity of STKs, which not only accounts for the conformational dynamics of the DFG-Phe residue but also encompasses the variability of other key structural motifs, namely the αC-helix and the activation loop, within the broader DFG-Phe ensemble.

A key challenge in sampling the conformational heterogeneity of apo STKs is that the timescales associated with transitions between metastable states are often too long to be captured by classical unbiased molecular dynamics (MD) simulations[8,9]. These transitions require navigating a rugged free energy landscape with numerous microstates, causing classical MD simulations to become trapped in local minima[10]. This issue is further exacerbated when studying how ligand binding to specific protein states modulates shift in populations and kinetics among functionally relevant metastable states (Figure 1). Traditional approaches are limited to specialized MD simulation hardware, such as Anton[11], and struggle to capture the slowly varying structural features of different protein domains that govern conformational shifts in biomolecules.

To address this sampling challenge, we utilized a previously developed framework[12] that employs multiple sequence alignment (MSA) subsampling[13,14] with AlphaFold to generate conformational heterogeneity. This approach uses multiple independent MD simulations from the AlphaFold generated conformational ensemble and Markov state model (MSM)[15–17] to sample the complete spectrum of conformational heterogeneity in apo STKs. However, this method can still become trapped in free energy basins if the initial structural diversity and subsequent MD simulations fail to overcome local free energy barriers.

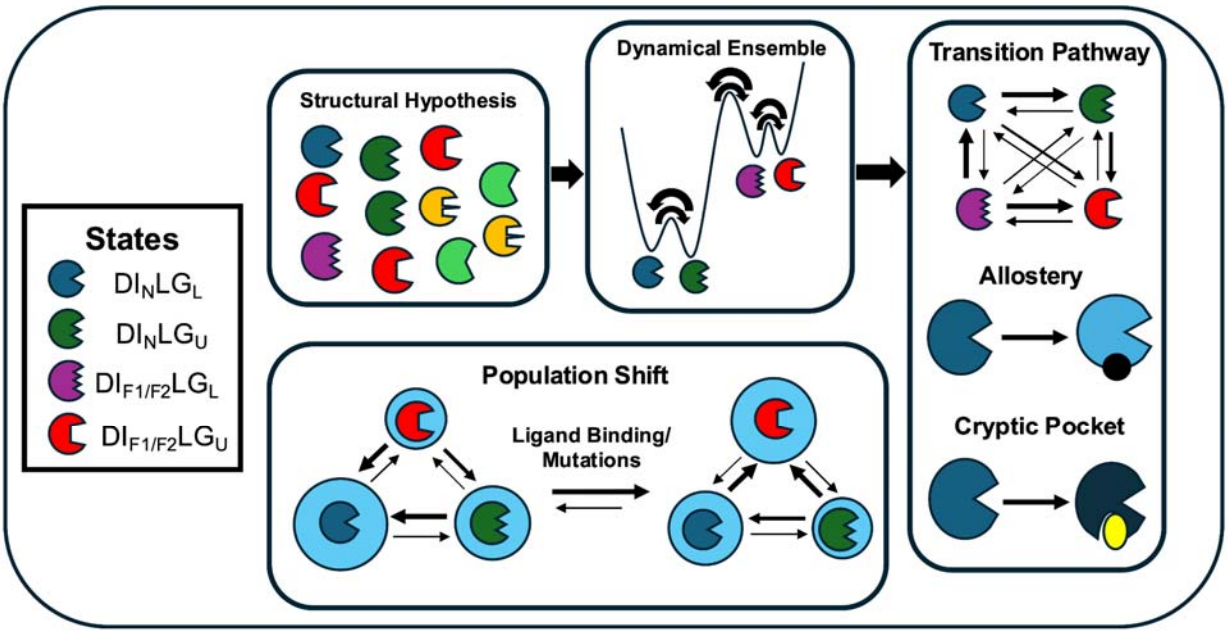

Figure 1. Our approach introduces an ensemble definition (states) of kinases, showing how AI/ML-augmented molecular simulations capture the conformational heterogeneity of kinase states and the effects of perturbations—such as mutations and ligand binding—on their populations and kinetics. While protein structure prediction models such as AlphaFold provide structural hypotheses, they lack the physics needed to accurately capture conformational dynamics. In contrast, physics-based molecular simulations informed by these structural hypotheses combined with machine learning reveal state populations, transition pathways, conformational allostery, and novel druggable cryptic pockets, expanding the scope of early-stage drug discovery.

To overcome this limitation, we developed a new adaptive sampling protocol[18] based on training a time-lagged autoencoder on short, physics-based molecular simulations seeded from AlphaFold generated ensembles. This protocol captures a physics-refined diverse conformational ensemble, enabling the sampling of protein conformations from underexplored regions of conformational space. Molecular simulations starting from this physics-refined conformational ensemble allowed us to capture a complete conformational spectrum of different functionally relevant metastable states in STKs. The weights from the *time-lagged* autoencoder[19], mapped to

protein structures, enabled us to identify structurally important conformational features that can guide the interpretation of experimental results and inform the design of new experiments. Further, the ability to capture physics refined conformational ensemble rather than single structure allowed us to capture conformational allostery[8,20,21] and cryptic pockets[12,22] in previously undruggable kinase-protein complexes.

The traditional research paradigm for kinases has primarily focused on retrospectively analyzing protein conformations[5] and sampling conformational transitions between known end states. In contrast, the ability to prospectively sample the conformational heterogeneity in a diverse set of apo STKs with different gatekeeper residues and ability to capture population shifts and the kinetics associated with ensemble perturbations, allows for a deeper understanding of mechanistic insights. Our frameworks can be extended across a wide range of kinase subclasses and other dynamic protein families, such as G protein-coupled receptors (GPCRs)[23] and tumor necrosis factors (TNFs)[24], to gain mechanistic insights and identify druggable pockets. This framework can de-risk and accelerate early-stage drug discovery across a wide-range of targets.

The ability of our pipeline to sample "*rare*" or "*out-of-distribution*" conformational ensembles in biomolecules can be exploited as a data generation tool. Our study shows that generating high-quality data that captures a spectrum of conformational heterogeneity in biomolecules is key to gain generalizable mechanistic insights of molecular motion in protein classes.

Results

Defining Conformational Heterogeneity in Serine-Threonine Kinases Beyond DFG In/Out

Traditional classification of protein kinase structures has focused on clustering kinase macrostates into **DFG-in** and **DFG-out** conformations. This classification is based on the dihedral angle of the DFG-Phe sidechain rotamer and distance vectors involving the DFG motif and the αC-helix (**Figure 2A**). However, these nomenclature schemes overlook the conformational heterogeneity within the DFG-in ensemble. Using our state-of-the-art computational pipeline, we comprehensively sampled the full range of conformational heterogeneity within the DFG-in ensemble (detailed in the following sections). This enabled us to introduce a novel nomenclature scheme that fully captures the observed conformational diversity in serine-threonine kinases.

Ensemble Definition

We categorized the global DFG-in (**DI**) ensemble into three distinct states—**DI_N**, **DI_{F1}**, and **DI_{F2}**—based on the χ angle (**Figure 2D**) of the conserved DFG-Phe side chain:

a. **DI_N**: χ angle between –2 radians and 0 radians.
b. **DI_{F1}**: χ angle between 0 radians and +2 radians.
c. **DI_{F2}**: χ angle below –2 radians or above +2 radians.

Lys-Glu Interaction in Serine-Threonine Kinases

Serine-threonine kinases are often characterized by a conserved salt-bridge interaction between a lysine residue in the β strand and a glutamate residue in the α-C helix (Figure 2A). To capture the heterogeneity of the Lys-Glu (**LG**) interaction, we introduce the following criteria:

1. **LG Latched (**LG_L**)**: Distance between Lys-Nζ and Glu-δ < 0.45 nm.
2. **LG Unlatched (**LG_U**)**: Distance between Lys-Nζ and Glu-δ > 0.45 nm.

The transition from **LG_L** to **LG_U** also reflects the α-C helix transitioning from an "in" to an "out" conformation (**Figure 2B**). The macrostates defined by **DI_N**, **DI_{F1}**, and **DI_{F2}** can each encompass multiple microstates, where the Lys-Glu interaction remains either latched or unlatched. These microstates are: **DI_NLG_L, DI_NLG_U, $DI_{F1}LG_L$, $DI_{F1}LG_U$, $DI_{F2}LG_L$, $DI_{F2}LG_U$** .

Activation loop dynamics in Serine-Threonine Kinases

The activation loop in serine/threonine kinases (STKs) plays a crucial role in molecular recognition and exists in a dynamic equilibrium between extended and collapsed conformations[25]. We capture the conformational heterogeneity of the activation loop by measuring the distance between a stable anchor point (DFG-Asp) and the conserved arginine residue within the activation loop (**Figure 2C**). The specific residue numbers for the anchor point and arginine are highlighted in the Supporting Information. This measurement allows us to define the conformational states as **AC_{out}** (extended) and **AC_{in}** (collapsed). When combined with the conformational states defined by **$DI_{N/F1/F2}$** and **$LG_{U/L}$** our approach effectively captures the full spectrum of conformational heterogeneity of the activation loop in STKs.

Our ensemble definitions can be extended to include DFG-out (**DO**) ensembles, while preserving all other microstate definitions associated with Lys–Glu (**LG**) interactions and the dynamics of activation loop (**AC**). Since STKs remain in the **DI** ensemble in the absence of a type-II inhibitor, we will focus our discussion on the conformational heterogeneity of the **DI** ensemble and how perturbations influence the thermodynamics and kinetics of various metastable states.

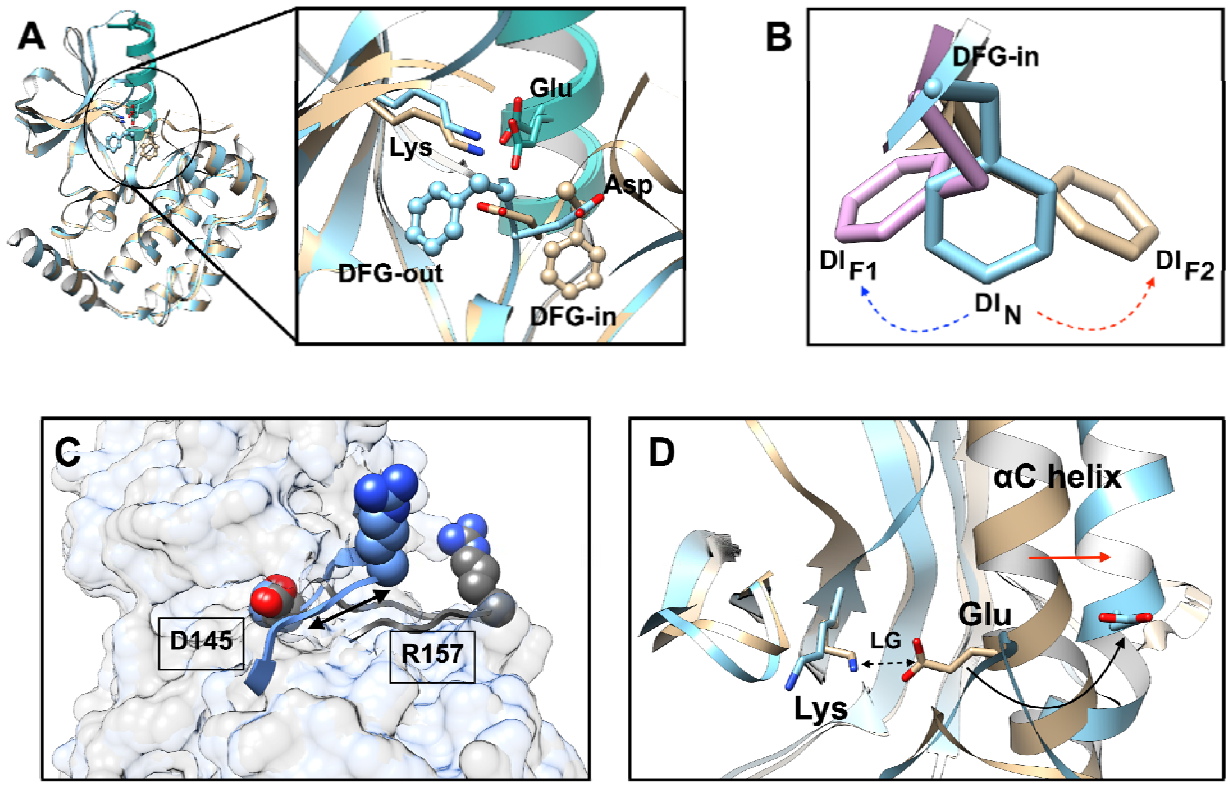

Figure 2. **A:** Pictorial representation of DFG-in/out snapshots in serine–threonine kinases, highlighting the positions of the catalytic DFG-Phe/Asp and conserved Lys and Glu residues. The αC-helix is shown in seagreen. **B:** Structural depiction of the $DI_{N/F1/F2}$ conformations within the DFG-in macrostate. **C:** The activation loop of CDK2 is highlighted as ribbon, and the distance (*C*□--*C*□) between Asp145 and Arg157 (indicated by the arrow) is used to define the extended (AC_{out}, grey) versus collapsed (AC_{in}, blue) conformations. **D:** Definition of latched (LG_L) versus unlatched (LG_U) conformations involving Lys and Glu, underscoring the inward-to-outward shift of the αC-helix and the resulting changes in the relative positions of Lys and Glu.

V600E mutation and Ligand Binding Modulate BRAF Conformational Dynamics and BRAF–MEK1 Interactions

The interaction between the serine-threonine kinase BRAF and MEK1 is a key step in the RAS–RAF–MEK–ERK signaling cascade, which regulates essential cellular processes such as growth, proliferation, and differentiation. BRAF is mutated in 66% of malignant melanomas, with the single-point mutation V600E accounting for approximately 80% of these cases. This BRAF V600E mutation results in a permanently active BRAF protein, continuously activating the MAPK pathway and driving cancer cell proliferation[26–28]. However, the specific effects of

V600E on the conformational dynamics of both the BRAF–MEK1 complex and the BRAF monomer remain an open question.

Physics-based molecular simulations of the wild-type BRAF–MEK1 complex, starting from an AlphaFold3-predicted structural ensemble, reveal that the complex predominantly adopts an inactive conformation, existing in a dynamic equilibrium between the $\mathbf{DI_{F1}}$ (78.71%) and $\mathbf{DI_{F2}}$ (21.29%) states (**Figure 3**). This inactive state also features Lys-Glu in unlatched conformations ($\mathbf{DI_{F1}LG_U}$, $\mathbf{DI_{F2}LG_U}$). In contrast, the V600E mutation shifts the population toward the active state (**Figure 4**, **Figure S2** in Supporting Information), evidenced by an increased population of the $\mathbf{DI_N}$ state (75.91%) compared to $\mathbf{DI_{F1}}$ (15.42%) and $\mathbf{DI_{F2}}$ (8.67%). Moreover, molecular dynamics simulations capture conformational heterogeneity within the $\mathbf{DI_N}$ state in V600E BRAF–MEK1, which transitions between $\mathbf{DI_NLG_L}$ and $\mathbf{DI_NLG_U}$ conformations (**Figure S1** in Supporting Information).

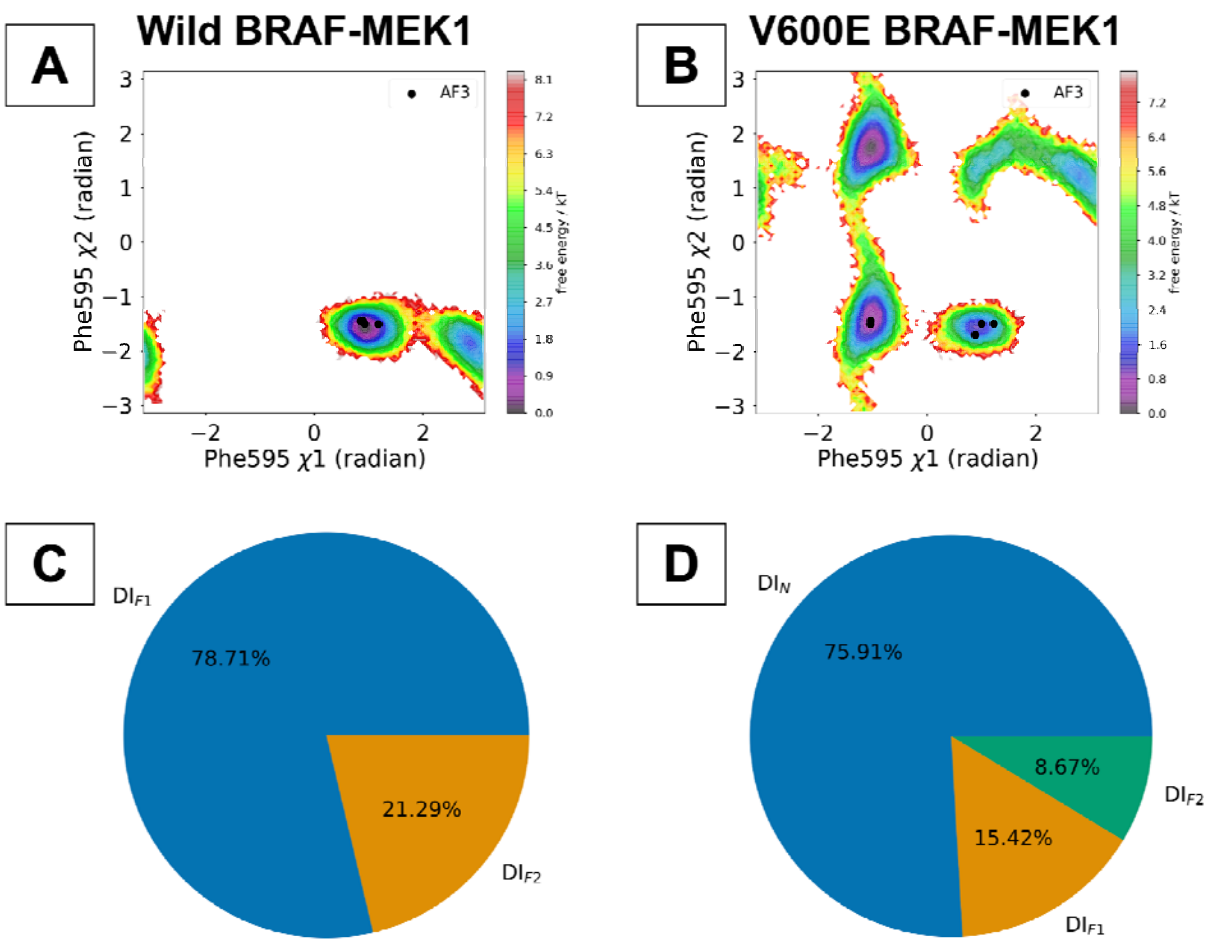

Figure 3. Markov state modeling (MSM) based on AI-augmented molecular simulations, initiated from an AlphaFold3-derived conformational ensemble, reveals conformational heterogeneity in the BRAF–MEK1 complex. In the wild-type complex, BRAF remains in a dynamic equilibrium between DI_{F1} and DI_{F2} states (A, C). However, the V600E mutation shifts the population from $DI_{F1/F2}$ to the DI_N state (B,

D). Although AlphaFold3 can sample single structures corresponding to DI_N and $DI_{F1/F2}$ conformations in V600E BRAF–MEK1, it does not capture the population shift upon mutation, emphasizing the importance of physics-based molecular simulations for a comprehensive characterization of conformational heterogeneity.

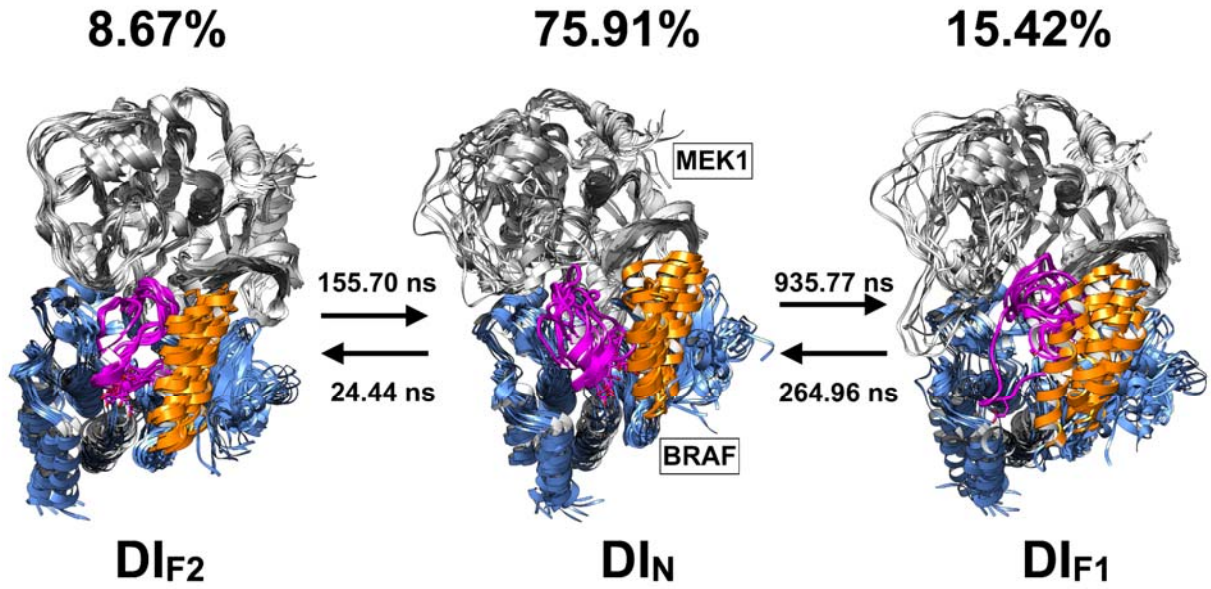

Figure 4. Timescales associated with conformational transitions in V600E BRAF–MEK1 complex reveal a faster transition from DI_{F1} to DI_N (264.96 ns) than from DI_N to DI_{F1} (935.77 ns). In contrast, wild-type BRAF–MEK1 remains in a dynamic equilibrium between DI_{F1} and DI_{F2}. Furthermore, the V600E BRAF–MEK1 complex exhibits conformational heterogeneity in the αC-helix (orange) and the activation loop (magenta), transitioning from an inward (DI_N) to outward (DI_{F1}) orientation in the αC-helix and from an extended (DI_N) to a collapsed (DI_{F1}) state in the activation loop.

Predicting the structural ensemble of the $\mathbf{DI_{F1}}$ conformation in the V600E BRAF–MEK1 complex enabled the identification of a hydrophobic backpocket in MEK1. Targeting this pocket with small molecules broadens the scope for designing novel therapeutic modalities that could stabilize the BRAF-MEK1 complex in the $\mathbf{DI_{F1}}$ conformation, opening a promising avenue for developing selective treatments against BRAF V600E–mutated melanoma (**Figure 5**).

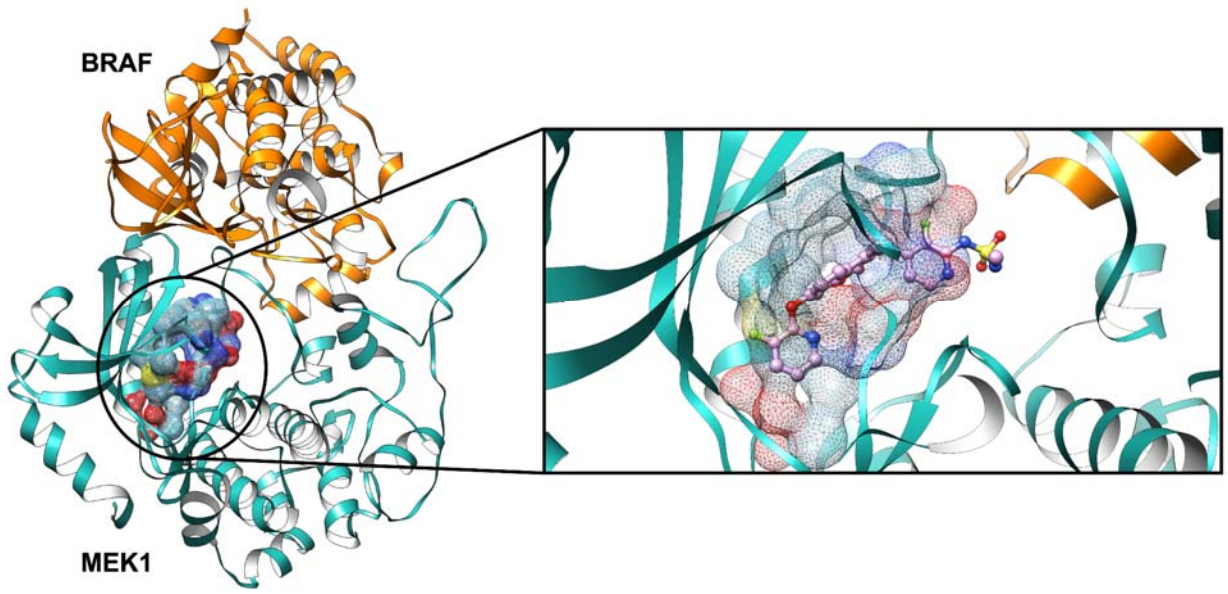

Figure 5. A structural snapshot from the conformational ensemble of the V600E BRAF–MEK1 complex reveals a hydrophobic backpocket in MEK1 with a druggability score of 0.880 predicted by FPocket[29]. Targeting this pocket with novel therapeutics can stabilize the V600E BRAF–MEK1 complex in the $\mathbf{DI_{F1}}$ state, potentially restoring normal cellular function. Notably, the panRAF–MEK1 molecular glue NST628 (PDB: 9AXX)[30], shown in ball-and-stick representation, binds to this hydrophobic backpocket as captured by our simulation.

Similar to the BRAF–MEK1 complex, the BRAF monomer also exhibits a population shift from the inactive ($\mathbf{DI_{F1/F2}}$) to the active ($\mathbf{DI_N}$) state upon V600E mutation (**Figure S4** in Supporting Information). Vemurafenib[31] was discovered as a highly specific, ATP-competitive inhibitor of BRAF V600E with selectivity against melanoma cells; however, how vemurafenib affects the conformational dynamics of V600E BRAF remains an open question. Molecular dynamics simulations initiated from V600E BRAF bound to vemurafenib in the $\mathbf{DI_N}$ state capture a shift in population towards the inactive state (**Figure 6** and **Figure S5** in Supporting Information), reflected by an increased population of the $\mathbf{DI_{F1}}$ and $\mathbf{DI_{F2}}$ states in the vemurafenib–V600E BRAF complex ($\mathbf{DI_{F1}}$: 23.50%, $\mathbf{DI_{F2}}$: 2.40%) compared to the apo V600E BRAF ($\mathbf{DI_{F1}}$: 12.86%, $\mathbf{DI_{F2}}$: 0.02%).

Conformational transitions from the $\mathbf{DI_N}$ state to the $\mathbf{DI_{F1}}$ state increase the population of $\mathbf{DI_{F1}LG_U}$ conformation in holo state (16.93%) compared to apo V600E BRAF (7.96%), a hallmark of active-to-inactive transitions in serine-threonine kinases (**Figure S1** in Supporting Information).

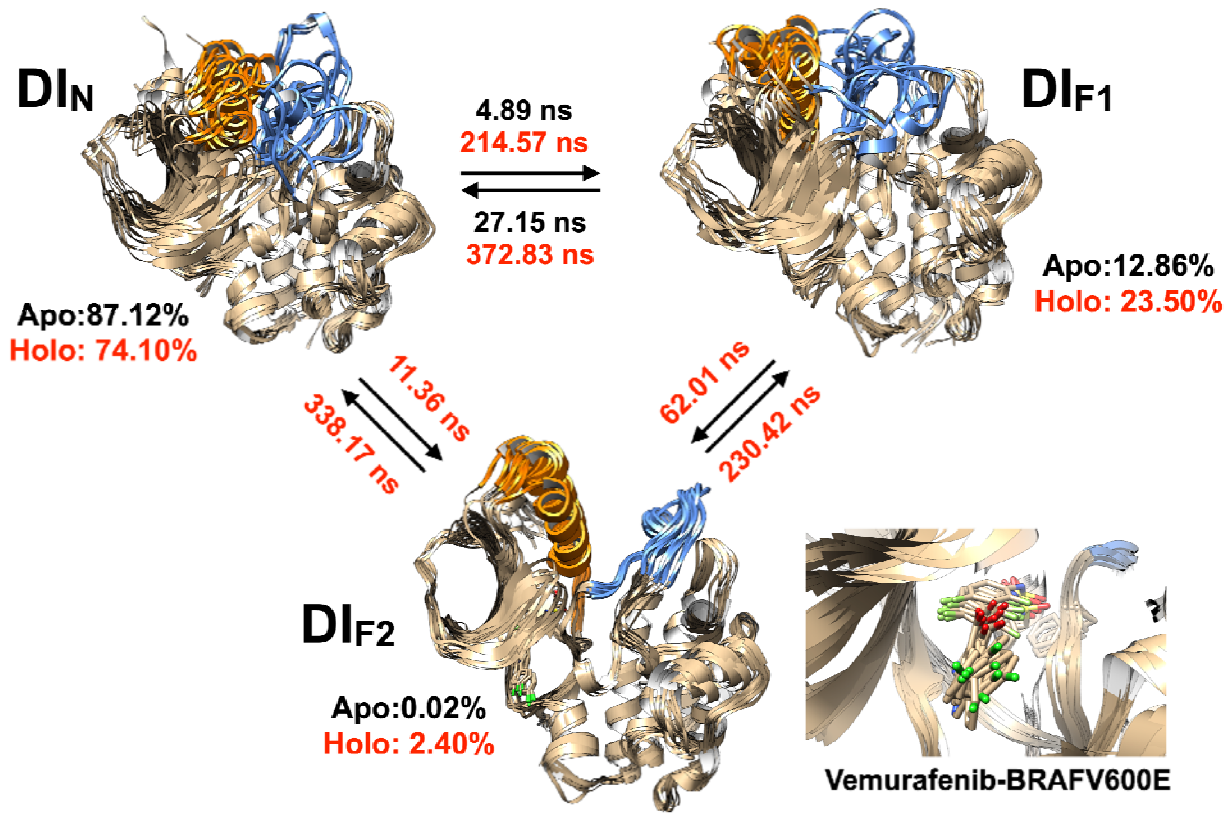

Figure 6. AI-augmented molecular simulations and ML-augmented adaptive sampling comprehensively capture the full range of conformational dynamics in the apo V600E BRAF monomer and the vemurafenib-bound holo V600E BRAF monomer respectively. Vemurafenib binding shifts the ensemble from DI_N to $DI_{F1/F2}$, as evidenced by faster $DI_N \rightarrow DI_{F1}$ (214.57 ns) and $DI_N \rightarrow DI_{F2}$ (11.36 ns) transitions compared to $DI_{F1} \rightarrow DI_N$ (372.83 ns) and $DI_{F2} \rightarrow DI_N$ (338.17 ns) in the holo form. Notably, the DI_{F2} state in apo V600E BRAF is sampled too sparsely to predict transition times via PCCA+[32]. The αC-helix and activation loop display marked conformational heterogeneity, highlighted in orange and blue, respectively.

Conformational Heterogeneity and Allostery in RIPK2: How Mutations and Protein-Protein Interactions Induce Population Shifts

The interaction between the kinase domain of receptor-interacting protein kinase 2 (RIPK2)—a serine-threonine kinase—and the E3 ligase X-linked Inhibitor of Apoptosis Protein (XIAP) serves as a crucial checkpoint in regulating inflammation and innate immunity, making RIPK2 a potential therapeutic target for diseases characterized by excessive or chronic inflammation[33–35]. A K47R mutation in RIPK2 alters its conformational state[36]. However, relying on a single static structure cannot adequately capture the dynamic nature of RIPK2, nor how mutations and protein–protein interactions modulate its conformational landscape.

To address this gap, a Markov state model (MSM) trained on AlphaFold-seeded molecular simulations was used to identify the populations and timescales associated with RIPK2's conformational states. In the wild-type protein, RIPK2 predominantly occupies the $\mathbf{DI_N}$ state (88.63%), with smaller populations in $\mathbf{DI_{F1}}$ (10.45%) and $\mathbf{DI_{F2}}$ (0.92%). Previously, $\mathbf{DI_N}$ was defined as the active state, while $\mathbf{DI_{F1}}$ and $\mathbf{DI_{F2}}$ were categorized as inactive states. In the K47R mutant, the conformational equilibrium is shifted toward the inactive states, with the population of $\mathbf{DI_{F1}}$ and $\mathbf{DI_{F2}}$ predicted to be 24.00% and 2.82% respectively (**Figure 7** and **Figure S7** in Supporting Information).

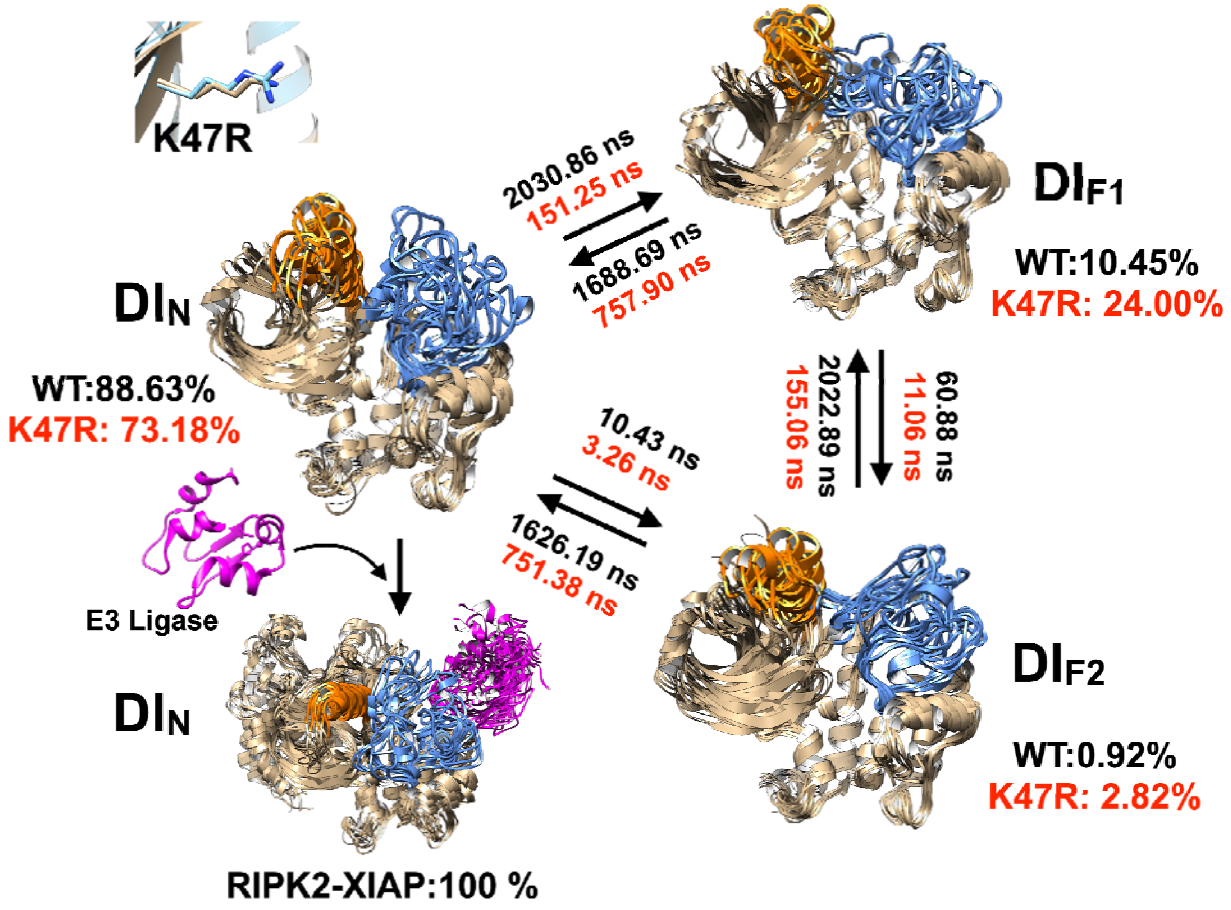

Figure 7. Populations and timescales associated with conformational transitions indicate that RIPK2 remains in a dynamic equilibrium among the $\mathbf{DI_N}$, $\mathbf{DI_{F1}}$, and $\mathbf{DI_{F2}}$ states. In wild type apo RIPK2, the active ($\mathbf{DI_N}$) state predominates. However, the K47R mutation shifts the equilibrium toward the inactive ($\mathbf{DI_{F1}}$ and $\mathbf{DI_{F2}}$) states. This population shift is reflected in the mean first-passage transition times, showing that the K47R mutation accelerates the $\mathbf{DI_N} \rightarrow \mathbf{DI_{F1/F2}}$ transitions compared with wild-type RIPK2. By contrast, binding of the E3 ligase XIAP stabilizes the $\mathbf{DI_N}$ state. The ability to predict conformational heterogeneity and understand how perturbations—such as mutations and XIAP binding—modulate RIPK2 dynamics is critical for capturing molecular recognition and developing selective therapeutics targeting *active-to-inactive* transitions. Conformational ensembles corresponding to each state capture the heterogeneity of the αC-helix (orange) and the activation loop (blue).

Additional physics-based molecular simulations also captured the conformational heterogeneity of microstates. In wild-type RIPK2, most of the population in the $\mathbf{DI_N}$ state remains in the Lys-Glu latched conformation ($\mathbf{LG_L}$) rather than the unlatched ($\mathbf{LG_U}$) conformation. Transitioning from $\mathbf{DI_N}$ to $\mathbf{DI_{F1}}$ shifts the population towards the unlatched conformation. This conformational change from $\mathbf{DI_N LG_L}$ to $\mathbf{DI_{F1} LG_U}$ involves a shift of the *αC-helix* from an "in" to an "out" conformation—a key hallmark of kinase conformational transitions (**Figure S8** in Supporting Information). Transition of *αC-helix* from in to out opens up the allosteric backpocket in RIPK2 which can be targeted by type III inhibitors[37] (**Figure 8**). It is key to highlight that AlphaFold alone failed to sample structures corresponding to $\mathbf{DI_{F1/F2}}$ conformations in wild type RIPK2.

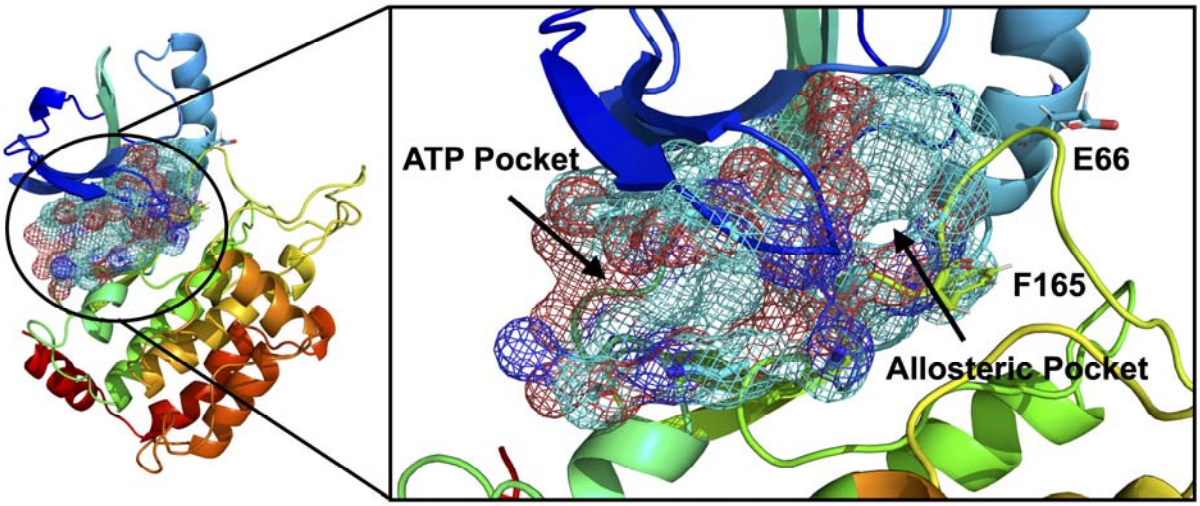

Figure 8. Pictorial representation of the $DI_{F1}LG_U$ conformational snapshot of apo RIPK2, illustrating the allosteric backpocket that can be targeted by type III kinase inhibitors. The formation of this backpocket is driven by the DI_N-to-DI_{F1} flip and the transition of the *αC-helix* from the "in" to "out" state.

Our time-lagged autoencoder based machine learning model (ML model) trained on high-dimensional temporal data from molecular dynamics simulations of RIPK2 identifies slowly varying structural features governing its conformational dynamics. The latent layers of the ML model serve as a two-dimensional collective variable—specifically, a nonlinear combination of dihedral angles—used to construct MSM (**Figure S9** and **S10** in Supporting Information). The model detects the flipping of Trp170 and Ile208. Flipping of Trp170 signifies a transition of the activation loop from an active to an inactive conformation as highlighted by *Vats et al*[21], whereas flipping of Ile208 alters the RIPK2–XIAP binding interface and disrupts the hydrophilic interaction involving Lys209 in RIPK2 and Asn209 and Glu211 in XIAP (**Figure 9**).

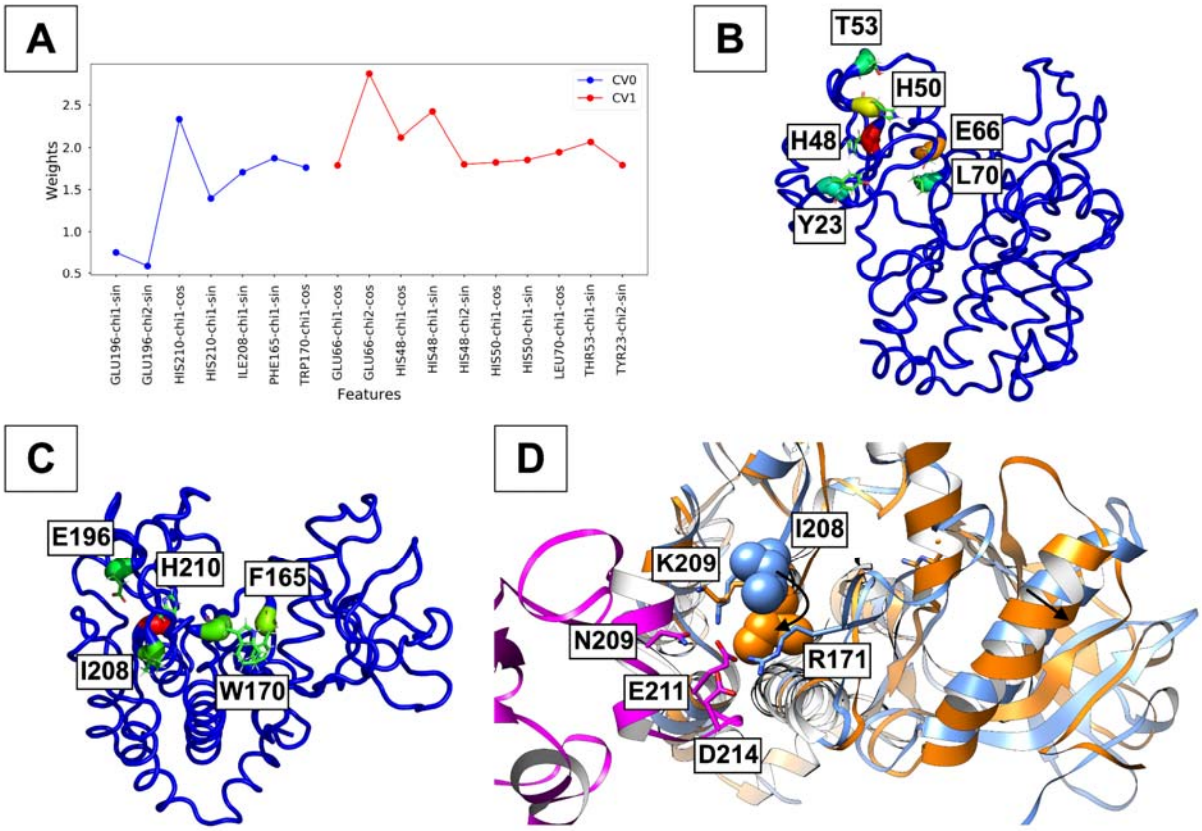

Figure 9. Latent layers of the ML model capture key structural features of RIPK2 (A). Specifically, CV0 reflects the C-lobe dynamics (defined by the χ1 and χ2 angles of residues 164–215), while CV1 encompasses the N-lobe dynamics (defined by the χ1 and χ2 angles of residues 14–74). ML weights projected onto the structural features reveal important conformational motions governing both N-lobe (B) and C-lobe (C) dynamics of RIPK2. In particular, the model highlights critical residues governing conformational heterogeneity of the αC-helix and the activation loop, which underpin transitions between RIPK2's active and inactive states. Within the activation loop, flipping of Trp170 disrupts the hydrogen bond between Arg171 of RIPK2 and Asp214 of XIAP. Simultaneously, flipping of Ile208 interrupts the interaction between Lys209 of RIPK2 and Asn209/Glu211 of XIAP (D). Together, these events illustrate the molecular basis of conformational allostery, showing how RIPK2 transitions from the active (blue) to the inactive (orange) state and how these changes modulate RIPK2–XIAP interactions.

Furthermore, MSM-predicted transition times indicate that the K47R mutation accelerates conformational transitions between the $\mathbf{DI_N} \leftrightarrow \mathbf{DI_{F1}}$ and $\mathbf{DI_N} \leftrightarrow \mathbf{DI_{F2}}$ states compared to wild-type RIPK2. Specifically, the mean first-passage time (MFPT) for the $\mathbf{DI_N} \rightarrow \mathbf{DI_{F1}}$ transition is ~2030 ns in the wild type and ~151 ns in the mutant, while the $\mathbf{DI_N} \rightarrow \mathbf{DI_{F2}}$ transition is ~10 ns in the wild type and ~3 ns in the mutant. A similar trend is observed for the reverse $\mathbf{DI_{F1/F2}} \rightarrow \mathbf{DI_N}$ transitions, highlighting that the K47R mutation shifts the population and the timescales of these

conformational transitions towards $\mathbf{DI_{F1/F2}}$ states which shifts the dynamics of the activation loop from extended ($\mathbf{AC_{out}}$) to collapsed ($\mathbf{AC_{in}}$) conformations (**Figure 7 and Figure S11 and S12 in Supporting Information**).

Unlike K47R mutation, RIPK2 and XIAP interaction stabilizes the $\mathbf{DI_N}$ state (**Figure 7** and **Figure S7** in Supporting Information). Taken together, these results demonstrate how both mutation and protein–protein interactions modulate the conformational dynamics of RIPK2, extending beyond the static structure paradigm of X-ray crystallography and AI-based predictions such as AlphaFold.

Conformational Dynamics of Serine-Threonine Kinases: A Generalizable Model

Markov state models (MSMs) enabled us to capture the populations and transition times of $\mathbf{DI_N}$, $\mathbf{DI_{F1}}$, and $\mathbf{DI_{F2}}$ states in an expanded set of serine-threonine kinases, including MAPK14[38], IRAK1 (Interleukin-1 receptor-associated kinase 1), IRAK4 (Interleukin-1 receptor-associated kinase 4)[39], and CDK2 (Cyclin-Dependent Kinase 2) [25]. In all of these proteins, the apo protein maintains a dynamic equilibrium among $\mathbf{DI_N}$, and $\mathbf{DI_{F1/F2}}$ conformations (**Figure 10**). Population-weighted projections along DFG-Phe $\chi1$ and Lys–Glu distances further revealed $\mathbf{LG_{L/U}}$ microstate heterogeneity within these $\mathbf{DI_{N/F1/F2}}$ macrostates (**Figure 11**).

Notably, in IRAK1, AlphaFold-seeded molecular dynamics simulations only partially sampled the conformational space defined by the $\mathbf{DI_{F1}}$, and $\mathbf{DI_{F2}}$ states. However, adaptive sampling initiated from conformational snapshots obtained through our ML model's latent layer captured the full spectrum of conformational heterogeneity (**Figure S13** in Supporting Information).

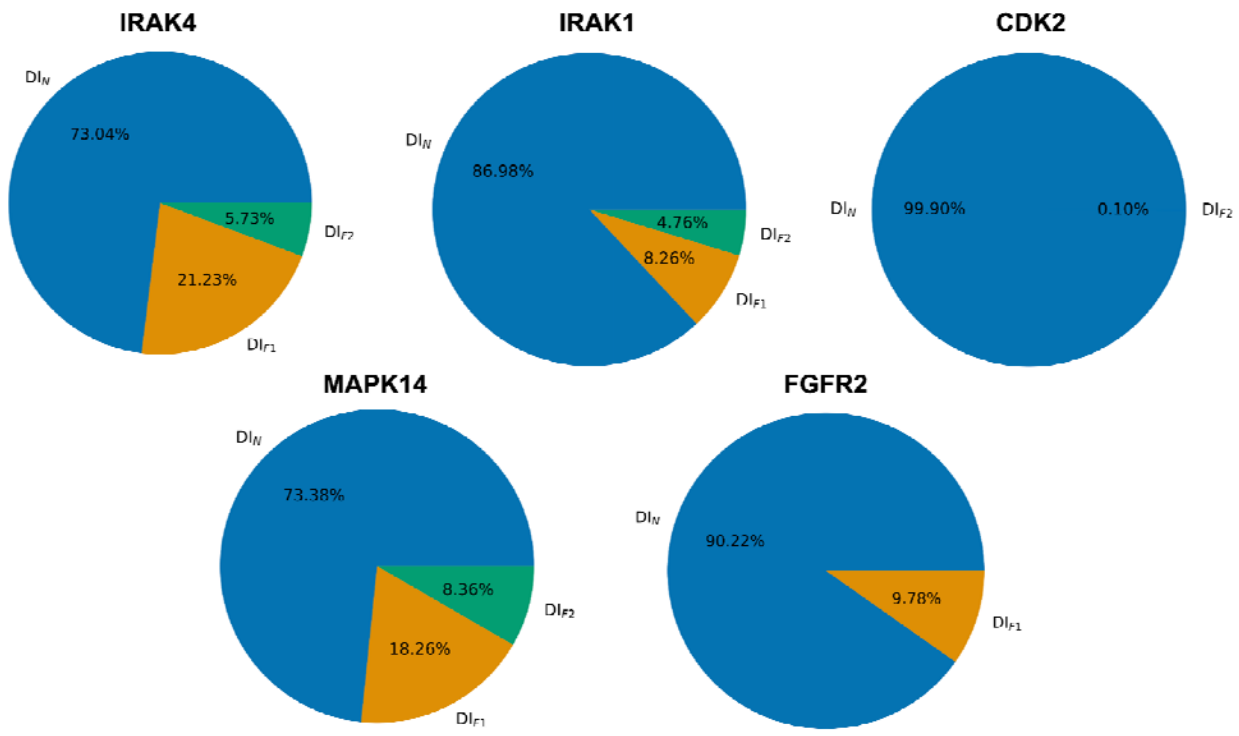

Figure 10. MSMs trained on physics-based molecular simulations of multiple serine-threonine kinases (IRAK1/4, CDK2, and MAPK14) and the tyrosine kinase FGFR2[40] reveals generalizable conformational dynamics, showing that kinases remain in dynamic equilibrium between DI_N and $DI_{F1/F2}$ states with variable populations.

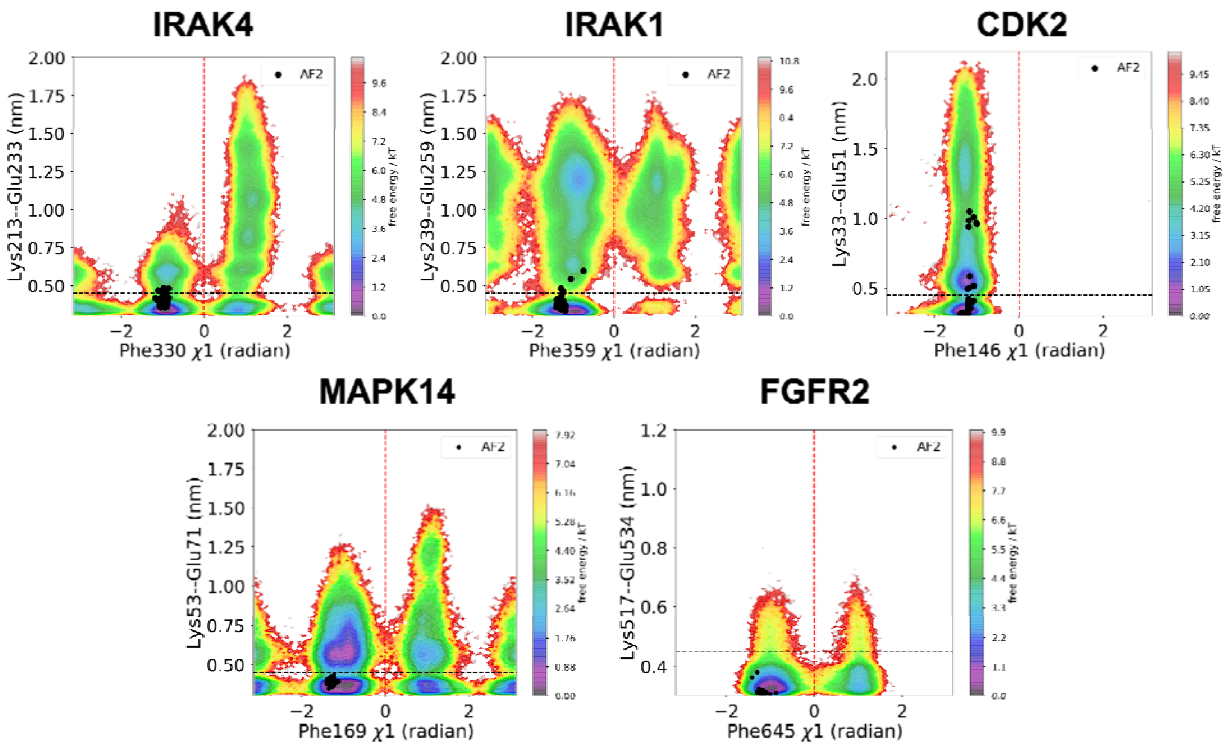

Figure 11. The population-weighted free energy surface derived from the MSM, projected along the χ1 angle of DFG-Phe and the Lys–Glu distance, reveals conformational heterogeneity defined by $\mathbf{LG_{L/U}}$ microstates within the broader $\mathbf{DI_{N/F1/F2}}$ macrostates. However, in all cases, conformational ensembles generated by MSA subsampling fail to capture the full spectrum of this heterogeneity.

Conformational variability in kinases extends beyond the $DI_{N/F1/F2}$ macrostates. For example, in the apo form of CDK2, the DI_N macrostate predominates but coexists in a dynamic equilibrium between AC_{out} (extended activation loop) and AC_{in} (collapsed activation loop) states involving the activation loop. Our ensemble definitions also capture conformational heterogeneity in the Lys-Glu interaction across $AC_{out/in}$ ensembles ($AC_{out/in}LG_{L/U}$). Cyclin E1 binding stabilizes the ACout conformation (**Figure 12** and **Figure S14** in the Supporting Information). Although Cyclin E1 has long been considered undruggable, a high-resolution dynamic representation of the CDK2–Cyclin E1 complex—encompassing its full conformational heterogeneity—reveals cryptic binding pockets within the protein–protein interface, opening new avenues for targeting Cyclin E1 (**Figure 13**).

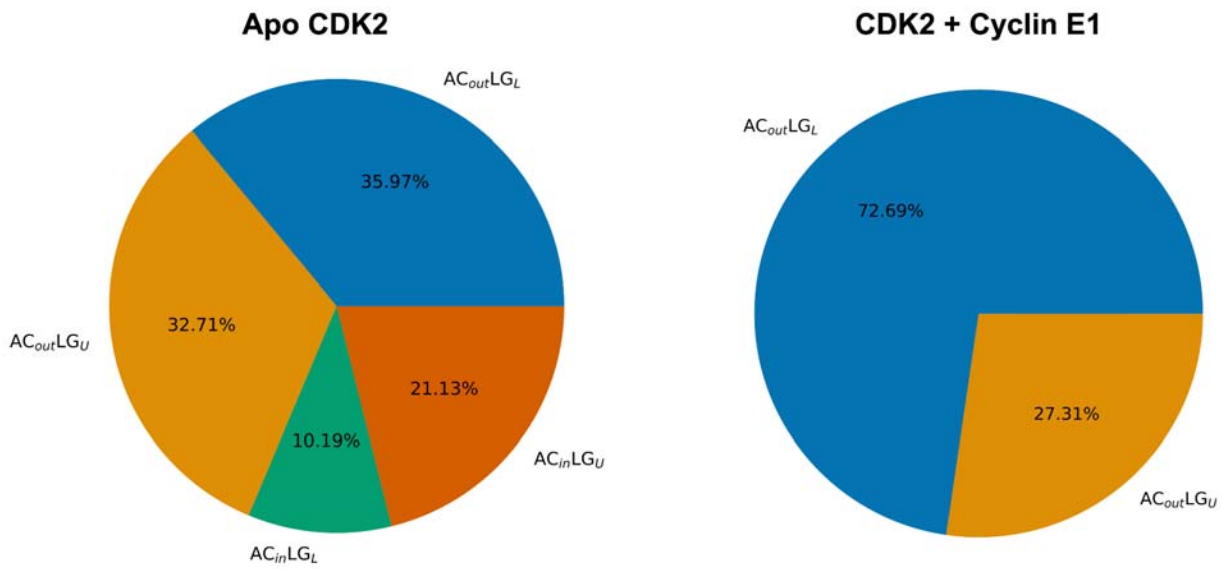

Figure 12. Our ensemble definition captures conformational heterogeneity of the activation loop and Lys-Glu interactions in apo CDK2 and CDK2-Cyclin E1 complex. Apo CDK2 remains in a dynamic equilibrium between AC_{out} and AC_{in} conformations which captures differentiable dynamics of the Lys-Glu interactions. Binding of Cyclin E1 stabilizes AC_{out} confirmation where the AC_{out} confirmation captures the dynamic equilibrium between LG_U and LG_L conformations.

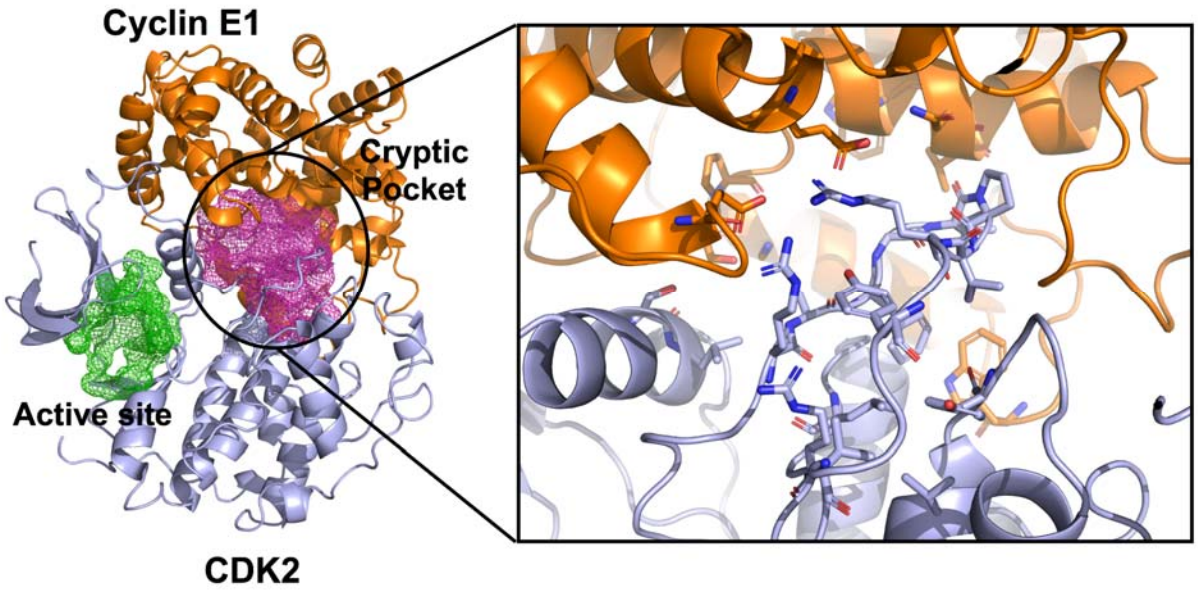

Figure 13. A structural snapshot from the DI_NAC_{out} conformational ensemble of the CDK2–Cyclin E1 complex reveals a cryptic pocket at the protein–protein interface with a high druggability score (0.967) predicted by FPocket. Targeting this cryptic pocket provides an alternative approach to modulating the CDK2–Cyclin E1 complex beyond the traditional CDK2 active site.

Finally, the conformational heterogeneity captured by our computational framework is not exclusive to serine-threonine kinases. Our adaptive sampling framework also enabled DI_N to DI_{F1} conformational transitions in receptor tyrosine kinase Fibroblast Growth Factor Receptor 2 (FGFR2) demonstrating the generalizability of our ensemble-definition method for sampling conformational dynamics across the human kinome (**Figure 14**). It is notable that ultra-long simulations initiated from the apo FGFR2 structure (PDB: 1GJO) on specialized Anton hardware, as well as conformational ensembles generated through MSA subsampling and AF seeded MD simulations (Figure S14 in Supporting Information), fail to capture $DI_N \rightarrow DI_{F1}$ conformational transitions (**Figure 14 C, D**). In contrast, binding of the highly selective small molecule RLY-4008[40] to FGFR2 stabilizes the DI_{F1} conformational state, underscoring the importance of prospectively capturing kinase conformational transitions for the development of selective therapeutics.

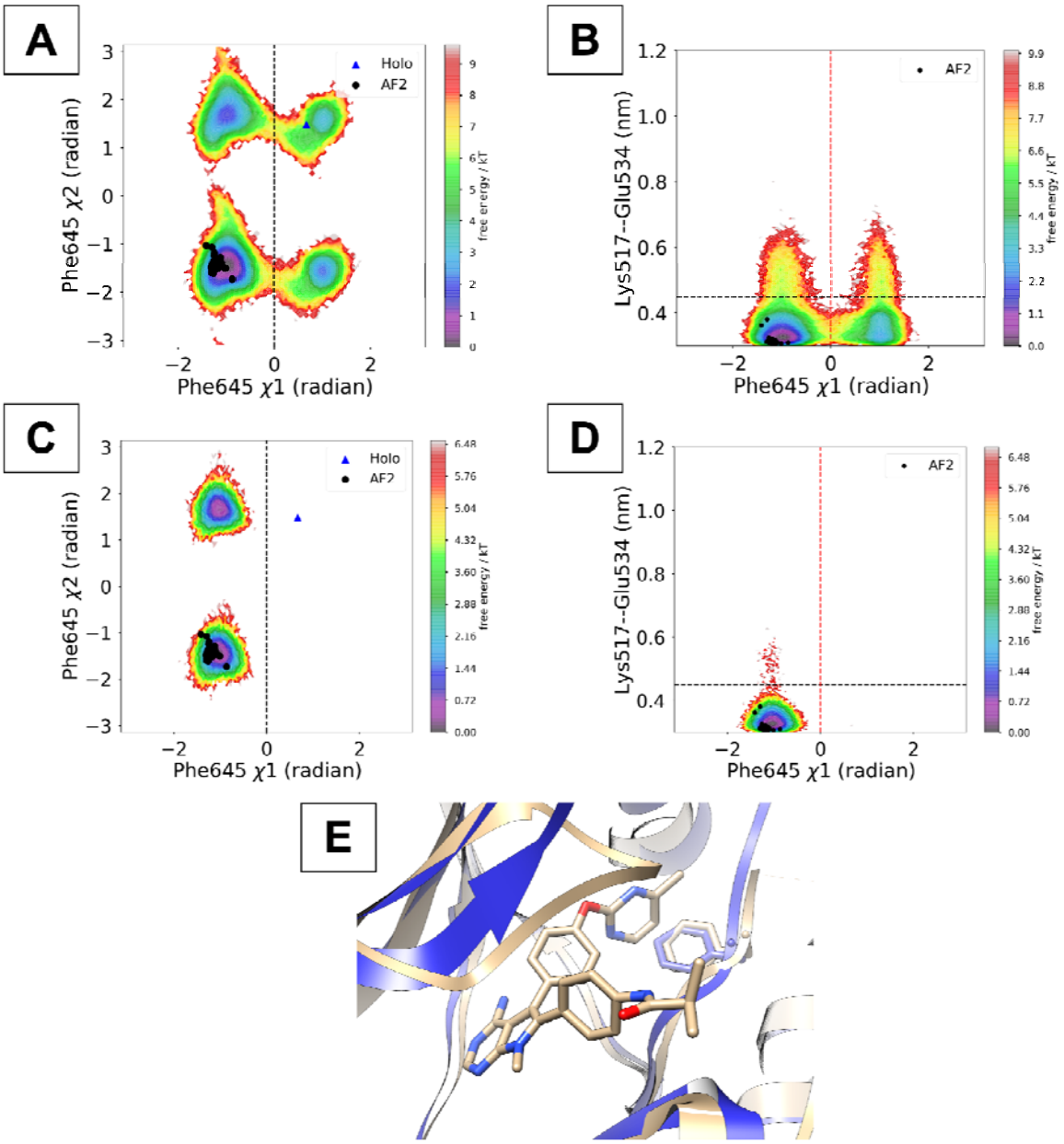

Figure 14. A population-weighted free energy surface, projected along DFG-Phe χ1/χ2 angles and the distance between DFG-Phe χ1 and Lys-Glu, illustrates how molecular simulation (100 replicas × 100 ns each, totaling 10 μs) initiated from physics-refined conformational ensemble captures apo FGFR2 conformational transitions (A, B) compared to ultralong MD simulations (three 25 μs replicas) performed on Anton. The blue triangle marks the holo FGFR2 structure (RLY-4008 + FGFR2 complex, PDB: 8STG[40]), while black dots represent conformational ensembles generated via MSA subsampling. A structural superimposition of the DI_{F1} state (blue) on holo FGFR2 (PDB: 8STG, silver) highlights the DFG-Phe conformation and the relative positioning of the small molecule (E).

Discussion

Serine/threonine kinases exist in a dynamic equilibrium among multiple metastable states. The conformational heterogeneity of these states governs their function and is critical for developing selective therapeutics. We introduced ensemble definitions to capture the full spectrum of conformational heterogeneity and employed AI-augmented molecular simulations and adaptive sampling to examine how the dynamical exchange among these ensembles is modulated by protein–protein interactions, small-molecule binding, and disease-causing mutations.

While the classic view of protein kinases has centered on the transition between DFG-in and DFG-out conformations—often induced by type-II inhibitors—this binary perspective does not fully capture the inherent conformational heterogeneity in apo kinases or how various perturbations shape their conformational ensembles. Our physics-based molecular simulations and Markov state models reveal that apo serine/threonine kinases (STKs) occupy a dynamic ensemble of three key macrostates—**DI_N**, **DI_{F1}**, and **DI_{F2}**—whose populations can be fine-tuned by different conformational perturbations. Notably, shifting from DI_N to DI_{F1} and DI_{F2}, where the DFG–Phe points toward the αC-helix, increases the fraction of broken Lys–Glu interactions, ultimately driving the αC-helix from its "*in*" to "*out*" conformation. Transitioning to $DI_{F1/F2}$ state together with αC-helix "out" conformation opens up allosteric backpocket which can be targeted by type-III small molecules to develop selective therapeutics against kinases[37,40].

Our simulations reveal that the V600E mutation in BRAF promotes a shift from the DI_{F1} to the DI_N population, activating the MAPK pathway and driving cancer cell proliferation. Vemurafenib, which binds to the hydrophobic back pocket of V600E BRAF, restores and stabilizes the DI_{F1} population. Likewise, other V600E-specific BRAF inhibitors—such as Dabrafenib and Encorafenib[41]—also stabilize the DI_{F1} state. Notably, our computational approach predicts that maintaining BRAF and MEK1 in the DI_{F1} conformation can restore normal cell division in V600E mutant cells. Building on this insight, we propose designing dual-target small molecules that bind the active sites of both BRAF and MEK1, shifting the population from DI_N to DI_{F1}—an approach that could serve as a combination therapy for treating V600E-driven cancers.

Beyond active-site inhibitors, designing novel classes of therapeutics offers an attractive strategy to reduce off-target toxicity and combat resistance. By capturing the DI_{F1} conformational ensemble in the V600E BRAF–MEK1 complex, we identified an expanded hydrophobic backpocket in MEK1. Small-molecule "glues" (such as NST-628) can target this site to stabilize the V600E BRAF–MEK1 complex in the DI_{F1} state, thereby restoring normal cellular function[30]. The ability to target an expanded hydrophobic backpocket in MEK1 and stabilize RAF and MEK1 in DI_{F1} state is an attractive strategy to develop pan-RAF-MEK1 molecular glue degraders.

Beyond small molecule binding selectivity, the ability to stabilize distinct conformational states is crucial for determining protein–protein interactions, which in turn govern key cellular functions and downstream signaling. For example, the E3 ligase XIAP binds to RIPK2 and stabilizes the D_{IN} state (PDB: 8AZA), whereas the interaction between BRAF and MEK1 selectively stabilizes the DI_{F1} state in normal cells. Small-molecule inhibitors of RIPK2, such as SB-203580 (PDB: 5AR4[42]), and CSLP18[43] (PDB: 6FU5), stabilize the DI_{F1} state, whereas molecules like K252 (PDB: 8X2O) favor the DI_N state. These observations suggest that DI_{F1} stabilizers act as PPI inhibitors, while DI_N stabilizers promote PPI interactions. Small molecules that stabilize the DI_{F1} conformation shift the equilibrium from the AC_{out} to the AC_{in} state, triggering a transition of the activation loop from an extended to a helical (collapsed) conformation (Figure S12 in Supporting Information). In the extended state, Arg171 forms a critical interaction with the E3 ligase XIAP. This extended-to-collapsed transition, along with an allosteric flip in Ile208 (**Figure 9**), modulates RIPK2–XIAP complex formation[21]. Because XIAP recruitment by RIPK2 is a key structural event in the NOD1/2 signaling pathway, our findings suggest that small molecules favoring the DI_{F1} state will exhibit excellent selectivity for RIPK2 and inhibit the RIPK2–XIAP interaction, thereby blocking NOD1/2 signaling.

This highlights the importance of understanding conformational heterogeneity and population shift in kinases to gain mechanistic insights into protein-protein interactions which modulate downstream signaling pathways implicated in inflammatory diseases. Our future study will combine machine learning augmented adaptive sampling and biochemical experiments to systematically characterize the effect of small molecules as on RIPK2-XIAP interactions, providing critical mechanistic insights into how small molecule binding allosterically modulates protein-protein interactions.

Recently, an enhanced sampling approach combining MSA subsampling, molecular dynamics (MD), and metadynamics with a variational autoencoder was applied to investigate the conformational transitions between DFG-in and DFG-out states[44]. In a follow up study, the authors applied this approach to kinase inhibitors[45], demonstrating how conformations selected from the DFG-in and DFG-out ensembles can be used for molecular docking. However, both these studies fail to capture how ligand binding modulates the conformational dynamics[46] within the DFG-in ensemble and allosterically modulates protein–protein interactions which governs downstream cellular signaling.

Here, we introduce a novel ML-augmented adaptive sampling protocol designed for two main purposes: (a) to capture functionally relevant slow conformational motions in different protein domains (**Figure S6** and **S13** in Supporting Information), revealing conformational allostery, and (b) to enable improved structural ensemble prediction through clustering in the latent layer of the ML model. In this scheme, short unbiased MD simulations are launched from the ensembles generated in the latent space, which, when combined with underlying physics, capture a more complete spectrum of conformational heterogeneity in both apo and holo kinases. This approach

goes beyond what can be achieved by AlphaFold-seeded simulations or traditional ultra-long MD.

A key feature of our protocol is the latent layer's ability to provide a compressed, two-dimensional representation of structural features originating from distinct high-dimensional inputs, such as the N-lobe and C-lobe dynamics of RIPK2. This separation of structural features is crucial for capturing slow conformational motions in larger proteins. We further demonstrate that our ML model captures the DI_N to DI_{F1} transition in RIPK2 and show how this transition allosterically influences the XIAP binding interface that governs RIPK2–XIAP complex formation. Additionally, we have highlighted how the latent layer of the ML model can capture '*slow*' conformational dynamics of the BRAF-MEK1 complex upon V600E mutation (**Figure S3** in Supporting Information). Finally, we highlight how the latent representations derived from the ML model can be employed as collective variables[47] for constructing MSMs (**Figure S9** and **S10** in Supporting Information) and for driving enhanced sampling simulations such as metadynamics.

Understanding how small-molecule binding modulates protein dynamics[48] without altering the underlying potential energy surface remains a major challenge, as conventional molecular simulations have limited ability to capture slow timescale conformational changes and sampled rugged free energy surface. Here, starting from the Vemurafenib-bound DI_N state of V600E BRAF, we demonstrate how one can sample $\mathbf{DI_N}$-to-$\mathbf{DI_{F1/F2}}$ conformational transitions using a multi-step protocol: (a) launch multiple independent short MD simulations, (b) train a machine learning (ML) model and generate conformational ensemble from its latent layer, (c) initiate new MD simulations from these latent-layer conformations, and (d) combine all simulations into a Markov state model (MSM). This workflow captures both the population shift and conformational heterogeneity of the holo V600E BRAF complex. Traditional structure-based drug discovery methods often rely on ranking small molecules solely by free energy values. However, consider a scenario where several compounds bind to a kinase with nearly identical free energy values, yet produce distinctly different effects on kinase-mediated protein–protein interactions. Conventional free energy approaches (e.g., FEP[49] or MM/GBSA[50]) struggle to capture these differences because they do not account for shifts in the conformational ensembles upon ligand binding[46]. Our adaptive sampling protocol directly addresses this limitation by revealing how small molecule binding modulates protein's conformational states. By illuminating transitions between distinct conformational states, this method explains how kinase-mediated protein–protein interactions can be modulated in different ways by ligands with similar free energy profiles. More broadly, the approach is applicable to any protein–ligand system and provides critical insights for designing highly selective therapeutics capable of differentially modulating protein–protein interactions through fine-tuning of protein dynamics.

Our ensemble definition extends beyond the $\mathbf{DI_{N/F1/F2}}$ states to include the conformational heterogeneity of the activation loop, as exemplified by CDK2. Although the DFG-Phe of CDK2

remains in the $\mathbf{DI_N}$ conformation, the activation loop remains in a dynamic equilibrium between collapsed ($\mathbf{AC_{in}}$) and extended ($\mathbf{AC_{out}}$) states. The activation loop dynamics, in turn, modulate the $\mathbf{LG_{L/U}}$ conformations. Understanding this interplay—particularly how the activation loop dynamics and the Lys-Glu interaction influence small-molecule binding or protein–protein interactions—is crucial for elucidating CDK2 signaling pathways. For instance, binding of selective small molecules to CDK2[51] (PDB: 8UV0, 9FR2, 7ZPC, 7SA0, 7S9X) shifts the conformational equilibrium toward the $\mathbf{AC_{in}LG_U}$ state, whereas CyclinE1 binding stabilizes the $\mathbf{AC_{out}LG_{U/L}}$ ensemble. Recognizing that CDK2's conformational dynamics extend beyond DFG-Phe is key to designing selective small molecules that induce the $\mathbf{AC_{in}}$ conformation or developing molecular glues that exploit the cryptic pocket at the CDK2–CyclinE1 interface. We propose that the conformational heterogeneity in CDKs is foundational to their interactions with cyclins and is key to develop selective therapeutics modulating specific CDK-Cyclin interactions.

Our ensemble definition can be extended to capture conformational heterogeneity in the DFG-out ensemble (**Figure S16** and **S17** in Supporting Information) and is not limited to serine/threonine kinases. Notably, our adaptive sampling framework prospectively sampled $\mathbf{DI_N}$ $_{\mathbf{to}}$ $\mathbf{DI_{F1}}$ $_{\mathbf{c}}$onformational transitions in the tyrosine kinase FGFR2 that was not detected by ultralong simulations by DE Shaw Research[40] initiated from the apo structure or traditional AlphaFold seeded molecular simulations (**Figure S15** in Supporting Information). The ability to prospectively sample conformational heterogeneity in kinases will be pivotal to integrate our accelerated computational pipeline with downstream traditional or AI augmented[52,53] population shift[54] based virtual screening pipeline to identify and develop selective therapeutics targeting a wide range of kinases. Furthermore, the ability to distinguish conformational populations across kinase homologs, as demonstrated for IRAK1/4 and FGFR1/2, can be generalized to other kinase families. This approach will enable the design of homologue-specific therapeutics with reduced off-target effects.

Conclusion

Our study underscores the importance of human interpretable ensemble definitions that capture not only endpoint structures but also the populations and timescales associated with conformational heterogeneity in protein subfamilies. However, much of the current work still focuses on rigidly defined conformational endpoints[13,14,55], overlooking the continuous spectrum of states and limiting our ability to understand how perturbations shift populations and timescales. The accelerated molecular simulation frameworks introduced here, combined with human interpretable ensemble definitions, can be extended to other protein subclasses—such as GPCRs, tumor necrosis factors (TNFs), intrinsically disordered proteins (IDPs)[56] —where conformational heterogeneity drives molecular recognition[57].

A major bottleneck to capturing rare yet functionally important conformational states—and thereby learning a generalized model of protein dynamics with interpretable collective variables—is the scarcity of training data. Our work emphasizes that high-quality conformational dynamics data is crucial for developing a generalizable understanding of protein dynamics. The recently proposed deep learning framework, *BioEmu*[58], has garnered attention for suggesting it may substitute molecular dynamics (MD) simulations. However, *BioEmu* cannot predict the thermodynamics and kinetics of population shifts under ensemble perturbations, nor is it capable of capturing generalized population shifts across different protein classes—primarily due to insufficient training data and inability to capture of '*out of distribution*' populations in the absence potential energy functions. In contrast, the computational framework described in our study can overcome these limitations by accurately modeling rare and functionally relevant conformational ensembles of biomolecules. The sampling approaches proposed in this study, combined with the diverse conformations available in the Protein Data Bank, can create high quality datasets needed to develop protein ensemble predictors from a given sequence within a specific protein subfamily. In the future, ensemble generation from ML model's latent layers, integrated with flow matching[59]/diffusion models[60] and Langevin dynamics[61], will provide a pathway to capture the full breadth of conformational heterogeneity across protein subfamilies—without requiring extensive molecular dynamics or enhanced sampling.

Methods

Structural ensemble generation using AlphaFold

A structural ensemble for each sequence of serine-threonine kinases (wild/K47R RIPK2, monomeric wild/V600E BRAF, CDK2, IRAK1, IRAK4, FGFR2, MAPK14) was generated using the ColabFold[62] implementation of AlphaFold2 (link: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=ADDuaolKmjGW), following the protocol outlined by Meller and co-workers[12]. We first built an initial multiple sequence alignment (MSA) via the MMseqs2[63] method integrated in ColabFold. Next, the MSA was subsampled stochastically to include up to 16 cluster centers and 32 additional sequences (referred to as *max_msa* = 16:32). For ensemble generation, we used the "*greedy*" pairing strategy to match any taxonomically compatible subsets. We used 16 random seeds and enabled model dropout. Leveraging both model dropout and multiple seeds capitalizes on AlphaFold2's[64] model uncertainties, resulting in 80 predicted structures per protein and capturing conformational heterogeneity.

Five protein–protein complexes, involving wild-type and V600E BRAF–MEK1 as well as CDK2–CyclinE1, were generated using AlphaFold3[65] (https://alphafoldserver.com/). The objective was to assess whether AlphaFold3 can capture the full spectrum of population shifts

and to determine if molecular simulations initiated from these AlphaFold3-generated conformations can reveal conformational heterogeneity.

Molecular dynamics simulations

Each apo protein structure was prepared for molecular dynamics simulations using the *tleap* module from Amber2022[66,67], following the protocols outlined by Meller et al.[12] and Vats et al.[21] Proteins were parameterized with the AMBER FF14SB[68] force field. For holo simulations, the small molecule (ligand) was parameterized with the GAFF[69] force field and partial charges were added using AM1-BCC[70] method. Counterions were added to each system, and the systems were then solvated in a truncated octahedron box containing TIP3P[71] waters, ensuring a minimum distance of 10 Å from the protein to the box boundary.

A two-phase energy minimization was performed on each system: a) Phase 1: Only the solvent and ions were minimized while restraining the protein with a 100 kcal/mol·Å² potential (200 steps of steepest descent followed by 200 steps of conjugate gradient), b) Phase 2: The entire system was minimized without restraints for 500 steps conjugate gradient.

Following minimization in Amber2022, the Amber topologies were converted to Gromacs format using Acpype[72]. Each system was then gradually heated from 0 K to 300 K over 500 ps in the NVT ensemble, applying harmonic restraints (500 kJ/mol·nm²) on the backbone of heavy atoms. Subsequently, the restraints were removed, and the systems were equilibrated for 200 ps in the NPT ensemble at 300 K, with pressure maintained at 1 bar by the Parrinello–Rahman barostat[73] and temperature controlled by the v-rescale thermostat.

Production simulations were conducted in the NPT ensemble at 300 K and 1 bar using the leapfrog integrator with a 2 fs timestep. A 1.0 nm cutoff was used for nonbonded interactions, while long-range electrostatic interactions were treated via the Particle Mesh Ewald (PME)[74] method with a 0.16 nm grid spacing. The LINCS[75] algorithm constrained covalent bonds to hydrogen atoms. All heating, equilibration, and production runs were carried out using GROMACS 2022[76].

For the AlphaFold-seeded molecular dynamics of monomeric RIPK2, CDK2, BRAF (wild and V600E), FGFR2, IRAK1/4, and MAPK14, a total of 80 structures were generated by AlphaFold. Each structure underwent a 100 ns simulation, culminating in 8 μs of total simulation time (80 × 100 ns).

Wild/V600E BRAF–MEK1 and CDK2-CylcinE1 complexes were generated using AlphaFold3 and prepared according to the aforementioned protocol. For each complex, five structural predictions were generated, and two independent 200 ns simulations (5 snapshots * 2 replicas * 200 ns each = 2000 ns) were performed for each prediction. Additionally, RIPK2+XIAP

simulations from *Vats et al*.[21] were used for post processing. In all cases, trajectories were saved every 10 ps.

Slow Feature Analysis

Slow Feature Analysis (SFA)[77] is a dimensionality reduction technique designed for high-dimensional temporal data. Its primary objective is to transform the *J*-dimensional input signal, □(□), into output signals □□(□) = □□(□(□)) using a set of nonlinear functions gk(c). These output signals are optimized to minimize

$$\Delta \square_{\square} = \langle {}_{k}{}^{2} \rangle$$

where $_k$ represents the temporal derivative of □□(□), and ⬚⬚⬚ denotes temporal averaging. By minimizing Δ□□, SFA extracts features that vary slowly over time.

Vats et al.[21] highlighted how SFA can capture linear combinations of slowly varying structural features from high-dimensional data generated from molecular dynamics (MD) simulations. In this study, we demonstrate how SFA can be applied to capture slowly varying conformational dynamics in kinases. Furthermore, we utilized SFA to develop the Markov state model[15,16] (Figure S10 in Supporting Information), which enabled us to predict the populations of functionally relevant conformations in kinases.

Time-lagged autoencoder

A time-lagged autoencoder (TAE)[19] is a specialized neural network architecture designed to learn compact representations of time-series data while explicitly leveraging temporal structure. This approach is particularly powerful in scenarios such as molecular dynamics (MD) simulations, where the goal is to capture slowly varying conformational dynamics of biomolecules from high-dimensional data. Identifying these long-timescale conformational transitions is crucial for understanding phenomena like conformational allostery in kinases.

Unlike standard autoencoders—which aim to reconstruct the same input data—time-lagged autoencoders are trained to predict $x_{t+\tau}$ from x_t, where τ is a user-defined lag time. By focusing on a non-zero τ, the TAE learns latent representations z_t that capture the slowly evolving features of the molecular system, which are relevant for predicting future states.
The TAE architecture comprises a) Encoder: Transforms the current system state x_t into a lower-dimensional representation z_t. b) Decoder: Reconstructs the time-lagged version of the system, $x_{t+\tau}$, from the latent representation z_t.

By pairing x_t and $x_{t+\tau}$, the model naturally emphasizes slowly varying dynamics, since the most prominent changes occurring over lag time τ are embedded in the latent space.

Latent Layer Representation: Let $x_t \in \mathbb{R}^d$ be the high-dimensional molecular state at time t. In the encoder, the latent layer is often formulated as:

$$z_t = \phi(W_1 x_t + b_1),$$

where:

- $W_1 \in \mathbb{R}^{(m \times d)}$ is the encoder's weight matrix.
- $b_1 \in \mathbb{R}^m$ is the bias vector.
- $\Phi(\cdot)$ is a nonlinear activation function .
- $z_t \in \mathbb{R}^m$ is the low-dimensional latent representation.

The decoder then reconstructs $x_{t+\tau}$ from z_t, typically using:

$$\hat{x}_{t+\tau} = W_2 z_t + b_2,$$

Where $W_2 \in \mathbb{R}^{(d \times m)}$ and $b_2 \in \mathbb{R}^d$. The training objective is to minimize the reconstruction error:

$$L = \Sigma_t \|x_{t+\tau} - \hat{x}_{t+\tau}\|^2,$$

which encourages z_t to preserve information about the slow degrees of freedom predictive of future conformations.

The TAE's learned latent space provides a compressed, nonlinear representation of input features that captures slowly varying conformational dynamics in molecular systems. Our machine learning framework adopts a dynamic feature-selection strategy by applying a learnable masked kernel through a ReLU activation, ensuring that only features with non-zero weights are retained. This yields two clear benefits: the retained features are easier to interpret, and they can be more effectively used as collective variables (CVs) in enhanced sampling simulations.

To give the latent space more structure, we extended the method with two modules: (a) We infused Gaussian noise into the output of the encoder before further processing it with the decoder, which as a result enforces better separated long lived states. (b) we expanded the training objective with the VAMP score[78] estimated in the latent space, thereby ordering the long-lived states by the slowest motions.

During training, the "*CVEstimator*" module orchestrates the entire process by balancing the reconstruction loss, the penalty for large, masked kernel weights, and the VAMP loss—each weighted appropriately—to optimize both interpretability and dynamical fidelity. This integrated

approach results in an end-to-end pipeline where feature selection, sparse representation learning, and dynamical analysis via VAMP collectively yield a robust and physically meaningful representation of the input data.

This setup enabled us to identify slowly evolving structural features in various protein domains. These nonlinear combinations of structural features serve as collective variables for constructing Markov state models (Figure S4 in Supporting Information) or can be incorporated into enhanced sampling frameworks. In our study, we used sine/cosine-transformed dihedral angles of different kinase domains as inputs, with two latent layers to capture compressed representations of each feature set.

Adaptive sampling

Our adaptive sampling method preserves the underlying dynamics by restarting MD trajectories at carefully chosen seeds, without introducing biasing forces. This approach addresses the sampling challenges that can arise from AlphaFold-seeded molecular dynamics simulations. The sampling pipeline follows these steps:

1. Starting conformations: Apo Simulations: Short (10 ns each) independent molecular dynamics (MD) simulations were launched from an AlphaFold-generated conformational ensemble (total 80 structures, training data: 10 ns * 80 = 800 ns) of apo IRAK1 and FGFR2 obtained through MSA subsampling. **Holo Simulations:** Small molecule was placed in the active site of the DIN conformational state of V600E BRAF. From this holo starting structure, 20 independent 10 ns MD simulations (10 ns * 20 = 200 ns) were initiated.

2. Time-Lagged Autoencoder: A time-lagged autoencoder was trained on the sin/cos-transformed $\chi1$ and $\chi2$ angles of the αC-helix and the activation loop in kinases (apo IRAK1, FGFR2 and the holo V600E BRAF monomer). This captures slowly varying conformational features from high-dimensional MD trajectory data (**Figure S6** and **S13** in Supporting Information).

3. Sampling conformations from the latent layers: *K-center*[79] clustering (K=100) was performed on the 2D latent space produced by the time-lagged autoencoder. Frames corresponding to each cluster center were extracted, representing the 'physics-refined' conformational ensemble.

4. Refined Simulations and Markov State Model: Independent 100 ns MD simulations were launched from each cluster center (total 100ns * 100 replicas = 10 microsecond) which were combined using the Markov state model to capture populations and timescales associated with conformational transitions.

Markov state model

Markov state models (MSMs) are a powerful framework for analyzing the conformational dynamics of biomolecules, providing a probabilistic description of the transitions between discrete states in a high-dimensional conformational landscape. MSMs leverage time-resolved data, such as that from molecular dynamics simulations, to model the system as a network of states connected by transition probabilities, capturing the likelihood of moving from one state to another over a defined lag time. Key steps in constructing an MSM include defining the conformational states—commonly achieved through clustering methods like k-means—and estimating the transition probabilities using a maximum-likelihood approach. By evaluating the dynamics at different lag times, MSMs enable the extraction of equilibrium populations and characteristic timescales of conformational transitions.

Markov state modeling (MSM) was performed on the sine and cosine transformations of the $\chi1$ and $\chi2$ angles of DFG-Phe, extracted from unbiased molecular dynamics simulations (launched from an AlphaFold-generated conformational ensemble) and from adaptive sampling simulations. K-means clustering (k = 100) was performed in the transformed dihedral space, and maximum-likelihood MSMs were generated using the lag times highlighted in the **Supporting Information**. Equilibrium populations obtained from these MSMs were projected along our ensemble definition to capture conformational heterogeneity in serine-threonine kinases. The convergence of MSM was assessed by predicting populations of different conformational states at different lag times (**Figure S9-S11, S2 and S4** in Supporting Information). PCCA+[32] was employed to define macrostates that distinguish conformational states associated with DFG-Phe. The lag time of 10 frames was used to build final MSMs unless explicitly mentioned.

For apo RIPK2, we used the 2D latent layer of a TAE and the first two slow features extracted from slow feature analysis, illustrating how one can build an MSM on a compressed representation. MSM generations were performed using PyEMMA 2.5.7[80].

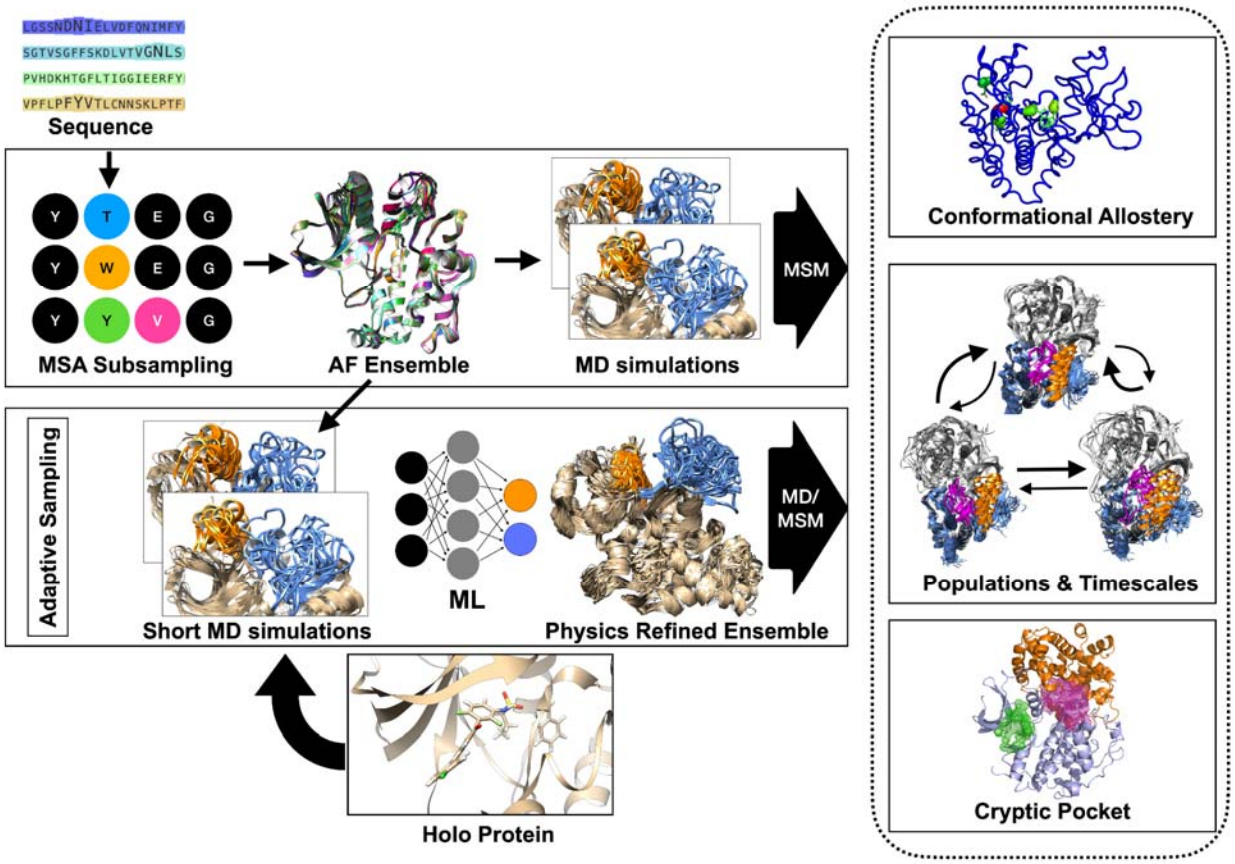

Figure 14. Overview of the computational framework employed in this study to capture generalizable dynamics in kinases and how perturbations such as ligand binding/mutations modulate conformational dynamics. The framework requires a protein sequence as input and uses the MSA subsampling approach to generate initial structural guesses (AF ensemble) using AlphaFold. These structures were then used to seed to launch independent MD simulations and finally combined together using the Markov state model (MSM) to capture conformational dynamics in kinases. In some cases, this standard workflow may not fully capture the breadth of conformational heterogeneity. To address such scenarios, we used an adaptive sampling framework which starts with launching short MD simulations (10 ns) from an AF generated conformational ensemble. The high-dimensional data from these simulations are used to train a time-lagged autoencoder (TAE), which identifies slowly varying structural features. Clustering in the TAE's latent space yields a "physics-refined" ensemble, and representative cluster centers are selected as starting points for further MD simulations. This expanded ensemble is then integrated into the MSM, providing a more comprehensive view of kinase conformational heterogeneity in both apo and holo states. Capturing these kinase dynamics facilitates the study of conformational allostery and reveals druggable cryptic pockets in protein–protein complexes, potentially accelerating the development of targeted therapies.

Data Availability

Release of trajectory data (transformed order parameters), Jupyter Notebooks and inference codes with fully automated pipeline are in preparation.

Due to large size of the trajectories, they are available for non-commercial use by contacting the corresponding author.

Supporting Information

Free energy surfaces, populations of conformational states at different lag times and ML weights for different protein kinases are available in the Supporting Information.

Acknowledgement

Computational simulations were performed using Tufts High Performance Computing Cluster and standalone Linux Clusters with RTX4090 graphics cards.

Competing interests

Authors declare no conflict of interests. Soumendranath Bhakat and Shray Vats are cofounders of AlloTec Bio, Inc.

References

(1) Johnson, J. L.; Yaron, T. M.; Huntsman, E. M.; Kerelsky, A.; Song, J.; Regev, A.; Lin, T.-Y.; Liberatore, K.; Cizin, D. M.; Cohen, B. M.; Vasan, N.; Ma, Y.; Krismer, K.; Robles, J. T.; van de Kooij, B.; van Vlimmeren, A. E.; Andrée-Busch, N.; Käufer, N. F.; Dorovkov, M. V; Ryazanov, A. G.; Takagi, Y.; Kastenhuber, E. R.; Goncalves, M. D.; Hopkins, B. D.; Elemento, O.; Taatjes, D. J.; Maucuer, A.; Yamashita, A.; Degterev, A.; Uduman, M.; Lu, J.; Landry, S. D.; Zhang, B.; Cossentino, I.; Linding, R.; Blenis, J.; Hornbeck, P. V; Turk, B. E.; Yaffe, M. B.; Cantley, L. C. An Atlas of Substrate Specificities for the Human Serine/Threonine Kinome. *Nature* **2023**, *613* (7945), 759–766. https://doi.org/10.1038/s41586-022-05575-3.

(2) Kugler, V.; Schwaighofer, S.; Feichtner, A.; Enzler, F.; Fleischmann, J.; Strich, S.; Schwarz, S.; Wilson, R.; Tschaikner, P.; Troppmair, J.; Sexl, V.; Meier, P.; Kaserer, T.; Stefan, E. Kinases in Motion: Impact of Protein and Small Molecule Interactions on Kinase Conformations. eLife Sciences Publications, Ltd 2024. https://doi.org/10.7554/elife.94755.2.

(3) Pellicena, P.; Kuriyan, J. Protein–Protein Interactions in the Allosteric Regulation of Protein Kinases. *Curr Opin Struct Biol* **2006**, *16* (6), 702–709. https://doi.org/https://doi.org/10.1016/j.sbi.2006.10.007.

(4) Treiber, D. K.; Shah, N. P. Ins and Outs of Kinase DFG Motifs. *Chem Biol* **2013**, *20* (6), 745–746. https://doi.org/https://doi.org/10.1016/j.chembiol.2013.06.001.

(5) Modi, V.; Dunbrack, R. L. Defining a New Nomenclature for the Structures of Active and Inactive Kinases. *Proceedings of the National Academy of Sciences* **2019**, *116* (14), 6818–6827. https://doi.org/10.1073/pnas.1814279116.

(6) Gizzio, J.; Thakur, A.; Haldane, A.; Levy, R. M. Evolutionary Divergence in the Conformational Landscapes of Tyrosine vs Serine/Threonine Kinases. *Elife* **2022**, *11*, e83368. https://doi.org/10.7554/eLife.83368.

(7) Vijayan, R. S. K.; He, P.; Modi, V.; Duong-Ly, K. C.; Ma, H.; Peterson, J. R.; Dunbrack, R. L. Jr.; Levy, R. M. Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. *J Med Chem* **2015**, *58* (1), 466–479. https://doi.org/10.1021/jm501603h.

(8) Kornev, A. P.; Taylor, S. S. Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci* **2015**, *40* (11), 628–647. https://doi.org/10.1016/j.tibs.2015.09.002.

(9) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.

(10) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced Sampling Techniques in Molecular Dynamics Simulations of Biological Systems. *Biochimica et Biophysica Acta (BBA) - General Subjects* **2015**, *1850* (5), 872–877. https://doi.org/https://doi.org/10.1016/j.bbagen.2014.10.019.

(11) Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. Long-Timescale Molecular Dynamics Simulations of Protein Structure and Function. *Curr Opin Struct Biol* **2009**, *19* (2), 120–127. https://doi.org/https://doi.org/10.1016/j.sbi.2009.03.004.

(12) Meller, A.; Bhakat, S.; Solieva, S.; R. Bowman, G. Accelerating Cryptic Pocket Discovery Using AlphaFold. *J Chem Theory Comput* **2023**, *19* (14), 4355–4363. https://doi.org/10.1021/acs.jctc.2c01189.

(13) del Alamo, D.; Sala, D.; Mchaourab, H. S.; Meiler, J. Sampling Alternative Conformational States of Transporters and Receptors with AlphaFold2. *Elife* **2022**, *11*, e75751. https://doi.org/10.7554/eLife.75751.

(14) Stein, R. A.; Mchaourab, H. S. SPEACH_AF: Sampling Protein Ensembles and Conformational Heterogeneity with Alphafold2. *PLoS Comput Biol* **2022**, *18* (8), e1010483.

(15) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J Am Chem Soc* **2018**, *140* (7), 2386–2396. https://doi.org/10.1021/jacs.7b12191.

(16) Pande, V. S.; Beauchamp, K.; Bowman, G. R. Everything You Wanted to Know about Markov State Models but Were Afraid to Ask. *Methods* **2010**, *52* (1), 99–105. https://doi.org/https://doi.org/10.1016/j.ymeth.2010.06.002.

(17) Konovalov, K. A.; Unarta, I. C.; Cao, S.; Goonetilleke, E. C.; Huang, X. Markov State Models to Study the Functional Dynamics of Proteins in the Wake of Machine Learning. *JACS Au* **2021**, *1* (9), 1330–1341. https://doi.org/10.1021/jacsau.1c00254.

(18) Kleiman, D. E.; Nadeem, H.; Shukla, D. Adaptive Sampling Methods for Molecular Dynamics in the Era of Machine Learning. *J Phys Chem B* **2023**, *127* (50), 10669–10681. https://doi.org/10.1021/acs.jpcb.3c04843.

(19) Wehmeyer, C.; Noé, F. Time-Lagged Autoencoders: Deep Learning of Slow Collective Variables for Molecular Kinetics. *J Chem Phys* **2018**, *148* (24), 241703. https://doi.org/10.1063/1.5011399.

(20) Guo, J.; Zhou, H.-X. Protein Allostery and Conformational Dynamics. *Chem Rev* **2016**, *116* (11), 6503–6515. https://doi.org/10.1021/acs.chemrev.5b00590.

(21) Vats, S.; Bobrovs, R.; Söderhjelm, P.; Bhakat, S. AlphaFold-SFA: Accelerated Sampling of Cryptic Pocket Opening, Protein-Ligand Binding and Allostery by AlphaFold, Slow Feature Analysis and Metadynamics. *PLoS One* **2024**, *19* (8), e0307226. https://doi.org/10.1371/journal.pone.0307226.

(22) Kuzmanic, A.; Bowman, G. R.; Juarez-Jimenez, J.; Michel, J.; Gervasio, F. L. Investigating Cryptic Binding Sites by Molecular Dynamics Simulations. *Acc Chem Res* **2020**, *53* (3), 654–661. https://doi.org/10.1021/acs.accounts.9b00613.

(23) Latorraca, N. R.; Venkatakrishnan, A. J.; Dror, R. O. GPCR Dynamics: Structures in Motion. *Chem Rev* **2017**, *117* (1), 139–155. https://doi.org/10.1021/acs.chemrev.6b00177.

(24) Croft, M.; Duan, W.; Choi, H.; Eun, S.-Y.; Madireddi, S.; Mehta, A. TNF Superfamily in Inflammatory Disease: Translating Basic Insights. *Trends Immunol* **2012**, *33* (3), 144–152. https://doi.org/10.1016/j.it.2011.10.004.

(25) Majumdar, A.; Burban, D. J.; Muretta, J. M.; Thompson, A. R.; Engel, T. A.; Rasmussen, D. M.; Subrahmanian, M. V; Veglia, G.; Thomas, D. D.; Levinson, N. M. Allostery Governs Cdk2 Activation and Differential Recognition of CDK Inhibitors. *Nat Chem Biol* **2021**, *17* (4), 456–464. https://doi.org/10.1038/s41589-020-00725-y.

(26) Kiel, C.; Benisty, H.; Lloréns-Rico, V.; Serrano, L. The Yin–Yang of Kinase Activation and Unfolding Explains the Peculiarity of Val600 in the Activation Segment of BRAF. *Elife* **2016**, *5*, e12814. https://doi.org/10.7554/eLife.12814.

(27) Cantwell-Dorris, E. R.; O'Leary, J. J.; Sheils, O. M. BRAFV600E: Implications for Carcinogenesis and Molecular Therapy. *Mol Cancer Ther* **2011**, *10* (3), 385–394. https://doi.org/10.1158/1535-7163.MCT-10-0799.

(28) Haling, J. R.; Sudhamsu, J.; Yen, I.; Sideris, S.; Sandoval, W.; Phung, W.; Bravo, B. J.; Giannetti, A. M.; Peck, A.; Masselot, A.; Morales, T.; Smith, D.; Brandhuber, B. J.; Hymowitz, S. G.; Malek, S. Structure of the BRAF-MEK Complex Reveals a Kinase Activity Independent Role for BRAF in MAPK Signaling. *Cancer Cell* **2014**, *26* (3), 402–413. https://doi.org/https://doi.org/10.1016/j.ccr.2014.07.007.

(29) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10* (1), 168. https://doi.org/10.1186/1471-2105-10-168.

(30) Ryan, M. B.; Quade, B.; Schenk, N.; Fang, Z.; Zingg, M.; Cohen, S. E.; Swalm, B. M.; Li, C.; Özen, A.; Ye, C.; Ritorto, M. S.; Huang, X.; Dar, A. C.; Han, Y.; Hoeflich, K. P.; Hale, M.; Hagel, M. The Pan-RAF–MEK Nondegrading Molecular Glue NST-628 Is a Potent and Brain-Penetrant Inhibitor of the RAS–MAPK Pathway with Activity across Diverse RAS- and RAF-Driven Cancers. *Cancer Discov* **2024**, *14* (7), 1190–1205. https://doi.org/10.1158/2159-8290.CD-24-0139.

(31) Nelson, B. E.; Roszik, J.; Janku, F.; Hong, D. S.; Kato, S.; Naing, A.; Piha-Paul, S.; Fu, S.; Tsimberidou, A.; Cabanillas, M.; Busaidy, N. L.; Javle, M.; Byers, L. A.; Heymach, J. V; Meric-Bernstam, F.; Subbiah, V. BRAF V600E–Mutant Cancers Treated with Vemurafenib Alone or in Combination with Everolimus, Sorafenib, or Crizotinib or with Paclitaxel and Carboplatin (VEM-PLUS) Study. *NPJ Precis Oncol* **2023**, *7* (1), 19. https://doi.org/10.1038/s41698-022-00341-0.

(32) Röblitz, S.; Weber, M. Fuzzy Spectral Clustering by PCCA+: Application to Markov State Models and Data Classification. *Adv Data Anal Classif* **2013**, *7* (2), 147–179. https://doi.org/10.1007/s11634-013-0134-6.

(33) Goncharov, T.; Hedayati, S.; Mulvihill, M. M.; Izrael-Tomasevic, A.; Zobel, K.; Jeet, S.; Fedorova, A. V; Eidenschenk, C.; deVoss, J.; Yu, K.; Shaw, A. S.; Kirkpatrick, D. S.; Fairbrother, W. J.; Deshayes, K.; Vucic, D. Disruption of XIAP-RIP2 Association Blocks NOD2-Mediated Inflammatory Signaling. *Mol Cell* **2018**, *69* (4), 551-565.e7. https://doi.org/10.1016/j.molcel.2018.01.016.

(34) Topal, J.; Gyrd-Hansen, M. RIPK2 NODs to XIAP and IBD. *Semin Cell Dev Biol* **2021**, *109*, 144–150. https://doi.org/https://doi.org/10.1016/j.semcdb.2020.07.001.

(35) Lethier, M.; Huard, K.; Hons, M.; Favier, A.; Brutscher, B.; Erba, E. B.; Abbott, D. W.; Cusack, S.; Pellegrini, E. Structure Shows That the BIR2 Domain of E3 Ligase XIAP Binds across the RIPK2 Kinase Dimer Interface. *Life Sci Alliance* **2023**, *6* (11), e202201784. https://doi.org/10.26508/lsa.202201784.

(36) Pellegrini, E.; Signor, L.; Singh, S.; Boeri Erba, E.; Cusack, S. Structures of the Inactive and Active States of RIP2 Kinase Inform on the Mechanism of Activation. *PLoS One* **2017**, *12* (5), e0177161.

(37) Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacol Res* **2016**, *103*, 26–48. https://doi.org/https://doi.org/10.1016/j.phrs.2015.10.021.

(38) Juyoux, P.; Galdadas, I.; Gobbo, D.; von Velsen, J.; Pelosse, M.; Tully, M.; Vadas, O.; Gervasio, F. L.; Pellegrini, E.; Bowler, M. W. Architecture of the MKK6-P38α Complex Defines the Basis of MAPK Specificity and Activation. *Science* **2023**, *381* (6663), 1217–1225. https://doi.org/10.1126/science.add7859.

(39) Bennett, J.; Starczynowski, D. T. IRAK1 and IRAK4 as Emerging Therapeutic Targets in Hematologic Malignancies. *Curr Opin Hematol* **2022**, *29* (1).

(40) Schönherr, H.; Ayaz, P.; Taylor, A. M.; Casaletto, J. B.; Touré, B. B.; Moustakas, D. T.; Hudson, B. M.; Valverde, R.; Zhao, S.; O'Hearn, P. J.; Foster, L.; Sharon, D. A.; Garfinkle, S.; Giordanetto, F.; Lescarbeau, A.; Kurukulasuriya, R.; Gerami-Moayed, N.; Maglic, D.; Bruderek, K.; Naik, G.; Gunaydin, H.; Mader, M. M.; Boezio, A. A.; McLean, T. H.; Chen, R.; Wang, Y.; Shaw, D. E.; Watters, J.; Bergstrom, D. A. Discovery of Lirafugratinib (RLY-4008), a Highly Selective Irreversible Small-Molecule Inhibitor of FGFR2. *Proceedings of the National Academy of Sciences* **2024**, *121* (6), e2317756121. https://doi.org/10.1073/pnas.2317756121.

(41) Gouda, M. A.; Subbiah, V. Expanding the Benefit: Dabrafenib/Trametinib as Tissue-Agnostic Therapy for BRAF V600E–Positive Adult and Pediatric Solid Tumors. *American Society of Clinical Oncology Educational Book* **2023**, No. 43, e404770. https://doi.org/10.1200/EDBK_404770.

(42) Charnley, A. K.; Convery, M. A.; Lakdawala Shah, A.; Jones, E.; Hardwicke, P.; Bridges, A.; Ouellette, M.; Totoritis, R.; Schwartz, B.; King, B. W.; Wisnoski, D. D.; Kang, J.; Eidam, P. M.; Votta, B. J.; Gough, P. J.; Marquis, R. W.; Bertin, J.; Casillas, L. Crystal Structures of Human RIP2 Kinase Catalytic Domain Complexed with ATP-Competitive Inhibitors: Foundations for Understanding Inhibitor Selectivity. *Bioorg Med Chem* **2015**, *23* (21), 7000–7006. https://doi.org/https://doi.org/10.1016/j.bmc.2015.09.038.

(43) Hrdinka, M.; Schlicher, L.; Dai, B.; Pinkas, D. M.; Bufton, J. C.; Picaud, S.; Ward, J. A.; Rogers, C.; Suebsuwong, C.; Nikhar, S.; Cuny, G. D.; Huber, K. V.; Filippakopoulos, P.; Bullock, A. N.; Degterev, A.; Gyrd-Hansen, M. Small Molecule Inhibitors Reveal an Indispensable Scaffolding Role of RIPK2 in NOD2 Signaling. *EMBO J* **2018**, *37* (17). https://doi.org/10.15252/embj.201899372.

(44) Vani, B. P.; Aranganathan, A.; Tiwary, P. Exploring Kinase Asp-Phe-Gly (DFG) Loop Conformational Stability with AlphaFold2-RAVE. *J Chem Inf Model* **2024**, *64* (7), 2789–2797. https://doi.org/10.1021/acs.jcim.3c01436.

(45) Gu, X.; Aranganathan, A.; Tiwary, P. Empowering AlphaFold2 for Protein Conformation Selective Drug Discovery with AlphaFold2-RAVE. eLife Sciences Publications, Ltd 2024. https://doi.org/10.7554/elife.99702.2.

(46) Olanders, G.; Testa, G.; Tibo, A.; Nittinger, E.; Tyrchan, C. Challenge for Deep Learning: Protein Structure Prediction of Ligand-Induced Conformational Changes at Allosteric and Orthosteric Sites. *J Chem Inf Model* **2024**, *64* (22), 8481–8494. https://doi.org/10.1021/acs.jcim.4c01475.

(47) Bhakat, S. Collective Variable Discovery in the Age of Machine Learning: Reality, Hype and Everything in Between. *RSC Adv* **2022**, *12* (38), 25010–25024. https://doi.org/10.1039/D2RA03660F.

(48) Khan, S. H.; Braet, S. M.; Koehler, S. J.; Elacqua, E.; Anand, G. S.; Okafor, C. D. Ligand-Induced Shifts in Conformational Ensembles That Describe Transcriptional Activation. *Elife* **2022**, *11*, e80140. https://doi.org/10.7554/eLife.80140.

(49) Abel, R.; Wang, L.; Harder, E. D.; Berne, B. J.; Friesner, R. A. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Acc Chem Res* **2017**, *50* (7), 1625–1632. https://doi.org/10.1021/acs.accounts.7b00083.

(50) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA Methods to Estimate Ligand-Binding Affinities. *Expert Opin Drug Discov* **2015**, *10* (5), 449–461. https://doi.org/10.1517/17460441.2015.1032936.

(51) Hummel, J. R.; Xiao, K.-J.; Yang, J. C.; Epling, L. B.; Mukai, K.; Ye, Q.; Xu, M.; Qian, D.; Huo, L.; Weber, M.; Roman, V.; Lo, Y.; Drake, K.; Stump, K.; Covington, M.; Kapilashrami, K.; Zhang, G.; Ye, M.; Diamond, S.; Yeleswaram, S.; Macarron, R.; Deller, M. C.; Wee, S.; Kim, S.; Wang, X.; Wu, L.; Yao, W. Discovery of (4-Pyrazolyl)-2-Aminopyrimidines as Potent and Selective Inhibitors of Cyclin-Dependent Kinase 2. *J Med Chem* **2024**, *67* (4), 3112–3126. https://doi.org/10.1021/acs.jmedchem.3c02287.

(52) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. *International Conference on Learning Representations (ICLR)* **2023**.

(53) Lu, W.; Zhang, J.; Huang, W.; Zhang, Z.; Jia, X.; Wang, Z.; Shi, L.; Li, C.; Wolynes, P. G.; Zheng, S. DynamicBind: Predicting Ligand-Specific Protein-Ligand Complex Structure with a Deep Equivariant Generative Model. *Nat Commun* **2024**, *15* (1), 1071. https://doi.org/10.1038/s41467-024-45461-2.

(54) Smith, L. G.; Novak, B.; Osato, M.; Mobley, D. L.; Bowman, G. R. PopShift: A Thermodynamically Sound Approach to Estimate Binding Free Energies by Accounting for Ligand-Induced Population Shifts from a Ligand-Free Markov State Model. *J Chem Theory Comput* **2024**, *20* (3), 1036–1050. https://doi.org/10.1021/acs.jctc.3c00870.

(55) Wayment-Steele, H. K.; Ojoawo, A.; Otten, R.; Apitz, J. M.; Pitsawong, W.; Hömberger, M.; Ovchinnikov, S.; Colwell, L.; Kern, D. Predicting Multiple Conformations via Sequence Clustering and AlphaFold2. *Nature* **2024**, *625* (7996), 832–839. https://doi.org/10.1038/s41586-023-06832-9.

(56) Lyle, N.; Das, R. K.; Pappu, R. V. A Quantitative Measure for Protein Conformational Heterogeneity. *J Chem Phys* **2013**, *139* (12), 121907. https://doi.org/10.1063/1.4812791.

(57) Bowman, G. R. AlphaFold and Protein Folding: Not Dead Yet! The Frontier Is Conformational Ensembles. *Annu Rev Biomed Data Sci* **2024**, *7* (1), 51–57. https://doi.org/10.1146/annurev-biodatasci-102423-011435.

(58) Lewis, S.; Hempel, T.; Jiménez-Luna, J.; Gastegger, M.; Xie, Y.; Foong, A. Y. K.; Satorras, V. G.; Abdin, O.; Veeling, B. S.; Zaporozhets, I.; Chen, Y.; Yang, S.; Schneuing, A.; Nigam, J.; Barbero, F.; Stimper, V.; Campbell, A.; Yim, J.; Lienen, M.; Shi, Y.; Zheng, S.; Schulz, H.; Munir, U.; Clementi, C.; Noé, F. Scalable Emulation of Protein Equilibrium Ensembles with Generative Deep Learning. *bioRxiv* **2024**, 2024.12.05.626885. https://doi.org/10.1101/2024.12.05.626885.

(59) Jing, B.; Berger, B.; Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles. *ArXiv* **2024**. https://doi.org/https://doi.org/10.48550/arXiv.2402.04845.

(60) Zheng, S.; He, J.; Liu, C.; Shi, Y.; Lu, Z.; Feng, W.; Ju, F.; Wang, J.; Zhu, J.; Min, Y.; Zhang, H.; Tang, S.; Hao, H.; Jin, P.; Chen, C.; Noé, F.; Liu, H.; Liu, T.-Y. Predicting Equilibrium Distributions for Molecular Systems with Deep Learning. *Nat Mach Intell* **2024**, *6* (5), 558–567. https://doi.org/10.1038/s42256-024-00837-3.

(61) Fan, J.; Li, Z.; Alcaide, E.; Ke, G.; Huang, H.; E, W. Accurate Conformation Sampling via Protein Structural Diffusion. *J Chem Inf Model* **2024**, *64* (22), 8414–8426. https://doi.org/10.1021/acs.jcim.4c00928.

(62) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6), 679–682. https://doi.org/10.1038/s41592-022-01488-1.

(63) Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat Biotechnol* **2017**, *35* (11), 1026–1028. https://doi.org/10.1038/nbt.3988.

(64) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(65) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(66) Case, D. A.; Aktulga, H. M.; Belfon, K.; Ben-Shalom, I. Y.; Berryman, J. T.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E. I.; Cisneros, G. A.; Cruzeiro, V. W. D.; Darden, T. A.; Forouzesh, N.; Giambaşu, G.; Giese, T.; Gilson, M. K.; Gohlke, H.; Goetz, A. W.; Harris, J.; Izadi, S.; Izmailov, S. A.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Machado, M.; Man, V.; Manathunga, M.; Merz, K. M.; Miao, Y.; Mikhailovskii, O.; Monard, G.; Nguyen, H.; O'Hearn, K. A.; Onufriev, A.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shajan, A.; Shen, J.; Simmerling, C. L.; Skrynnikov, N. R.; Smith, J.; Swails, J.; Walker, R. C.; Wang, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Xiong, Y.; Xue, Y.; York, D. M.; Zhao, S.; Zhu, Q.; Kollman, P. A. Amber 2022, University of California, San Francisco. **2022**.

(67) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *WIREs Computational Molecular Science* **2013**, *3* (2), 198–210. https://doi.org/https://doi.org/10.1002/wcms.1121.

(68) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255.

(69) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J Comput Chem* **2004**, *25* (9), 1157–1174. https://doi.org/https://doi.org/10.1002/jcc.20035.

(70) Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. *J Comput Chem* **2000**, *21* (2), 132–146. https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(20000130)21:2<132::AID-JCC5>3.0.CO;2-P.

(71) 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* **1983**, *79* (2), 926–935. https://doi.org/10.1063/1.445869.

(72) Sousa da Silva, A. W.; Vranken, W. F. ACPYPE - AnteChamber PYthon Parser InterfacE. *BMC Res Notes* **2012**, *5* (1), 367. https://doi.org/10.1186/1756-0500-5-367.

(73) Parrinello, M.; Rahman, A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J Appl Phys* **1981**, *52* (12), 7182–7190. https://doi.org/10.1063/1.328693.

(74) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N·log(N) Method for Ewald Sums in Large Systems. *J Chem Phys* **1993**, *98* (12), 10089–10092. https://doi.org/10.1063/1.464397.

(75) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem* **1997**, *18* (12), 1463–1472. https://doi.org/https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

(76) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1–2*, 19–25. https://doi.org/https://doi.org/10.1016/j.softx.2015.06.001.

(77) Wiskott, L.; Sejnowski, T. J. Slow Feature Analysis: Unsupervised Learning of Invariances. *Neural Comput* **2002**, *14* (4), 715–770. https://doi.org/10.1162/089976602317318938.

(78) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for Deep Learning of Molecular Kinetics. *Nat Commun* **2018**, *9* (1), 5. https://doi.org/10.1038/s41467-017-02388-1.

(79) Gonzalez, T. F. Clustering to Minimize the Maximum Intercluster Distance. *Theor Comput Sci* **1985**, *38*, 293–306. https://doi.org/https://doi.org/10.1016/0304-3975(85)90224-5.

(80) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J Chem Theory Comput* **2015**, *11* (11), 5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.