Geometry-Constrained Prediction of Catalytically Competent Kinase Domains Across the Human Kinome

Yuxuan Wang [1,2,3], Qi Hu [1,2,3], Jing Huang [1,2,3,*]
1 State Key Laboratory of Gene Expression, School of Life Sciences, Westlake University, Hangzhou 310030, Zhejiang, China
2 Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310024, Zhejiang, China
3 Institute of Biology, Westlake Institute for Advanced Study, Hangzhou 310024, Zhejiang, China
*Correspondence: huangjing@westlake.edu.cn (J.H.)

Abstract
Understanding cellular signaling networks benefits from accurate structural modeling of protein kinases across their different functional states. However, most human kinases lack experimentally determined active-state structures, and existing computational methods often fail to reliably predict catalytically competent conformations. Here, we present a geometry-constrained modeling framework based on AFEXplorer (AFEX) that enables robust prediction of active, ATP- and magnesium-bound states for 436 catalytic kinase domains across the human kinome. By explicitly enforcing key biophysical criteria, AFEX overcomes the template dependence and conformational bias inherent in AlphaFold2 and AlphaFold3. Benchmarking against 156 kinase-substrate complexes shows that AFEX achieves sub-2 Å backbone RMSD in 99% of cases and accurately recapitulates critical catalytic features with precise ATP coordination. Notably, AFEX correctly predicted active conformations for EGFR and CDK2, where AlphaFold3 failed. These improvements are consistent across diverse kinase families, including both well-characterized kinases and those underrepresented or absent from AlphaFold's training data. AFEX bridges a critical gap in kinase biology by providing a generalizable framework for studying active conformations and dynamics in kinases and other signaling proteins.

Introduction

Protein kinases are central regulators of intracellular phosphorylation signaling, orchestrating over 30% of cellular protein activity and governing critical processes such as proliferation, differentiation, and apoptosis [1]. The human kinome comprises 481 genes, each encoding at least one canonical full-length kinase domain; 13 of these genes contain two kinase domains, resulting in a total of 494 kinase domains [2,3]. Of these, 437 are predicted to be catalytically active and mediate the phosphorylation of serine, threonine, or tyrosine residues to modulate signaling networks [4]. These kinases are classified into eight major families (AGC, CAMK, CK1, CMGC, STE, TK, TKL, and RGC), each comprising distinct subfamilies with specialized roles in processes ranging from transcriptional regulation to cell cycle control [3]. All kinases share a conserved catalytic core, which includes an N-terminal β-strand lobe with a critical αC-helix and a C-terminal α-helical lobe. Crucially, catalytic activity depends on precise and uniform geometric arrangements: the activation loop must adopt an extended conformation, and the αC-helix must rotate inward to align the ATP-binding and substrate recognition sites [5,6]. In contrast, inactive states exhibit structural heterogeneity, such as DFG motif rearrangements or αC-helix displacement, which alter the accessibility of the binding pocket and impair catalytic function [6].

Structural characterization of active-state conformations across the catalytic human kinome is crucial for elucidating the biochemical mechanisms underlying kinase signaling. Conformational dynamics regulate catalytic activity, substrate recognition, allosteric control, and subcellular localization, collectively coordinating cellular responses to extracellular stimuli. In general, kinases act as molecular switches, cycling between active and inactive states through conformational changes in key structural elements such as the activation segment, αC-helix, and DFG motif. Disruptions to these regions, induced by mutations or inhibitors, can lead to aberrant signaling and are implicated in diseases such as cancer, inflammation, and neurodegeneration [7]. For example, NMR studies of Abelson kinase (Abl) demonstrate that the active conformation predominates (~88%) under physiological conditions, and drug-resistant mutations shift the conformational equilibrium to further stabilize the active state [8].

The transition into the active-state conformation is sometimes coupled to the relocation of kinases and changes in their binding partners, yet the atomistic basis and mechanistic understanding of these processes remain poorly defined. For example, activation of Protein Kinase D (PKD) is essential for the fission of transport carriers from the trans-Golgi network (TGN) to the plasma membrane. Kinase-inactive PKD-K618N mutants accumulate at the TGN, resulting in extensive membrane tubulation and defective vesicle release [9,10]. Activation of ERK1 triggers its nuclear translocation, marking the final step of the MAPK signaling cascade, where it engages transcription factors to regulate cell cycle progression [11]. Such activation-induced subcellular relocalization is a conserved regulatory mechanism: during apoptosis, PKCδ translocates to the nucleus to promote lamin B1 disassembly [12], whereas PKCε enhances ERK nuclear import [13], highlighting the functional significance of active-state conformations. Accurately capturing these active-state structures can therefore help elucidate the spatial and functional regulation of kinase signaling.

Despite the conservation of core activation motifs across the kinome, structural diversity precludes straightforward generalization. A major challenge in characterizing active-state conformations lies in the limited availability of experimentally resolved active kinase structures. Among 272 unique kinase domains deposited in the Protein Data Bank (PDB), only 156 meet established criteria for active states, and many lack complete resolution of the activation loop [4]. This paucity stems in part from inherent limitations of traditional structural biology techniques, such as X-ray crystallography and cryo-electron microscopy, which face challenges in sample preparation and data collection when resolving catalytically active conformations, particularly for low molecular weight kinases [14]. Deep learning–based methods, notably AlphaFold, have transformed protein structure prediction [15–17], yet due to biases in training data, these models often fail to capture functionally relevant

conformational states, including the active conformations of kinases. For instance, only 48% of catalytic domains in the EBI dataset exhibit active-like geometry [17]. Although AlphaFold-predicted structures have shown promise in facilitating ligand discovery [18], their utility is constrained by structural uncertainties in low-confidence regions, inaccuracies in side-chain modeling, and substantial variations in binding pocket volumes [19–21], underscoring the need for accurate, functionally relevant structural models.

Methods that integrate physics-based enhanced sampling techniques to generate diverse conformations have been proposed to improve the efficiency and selectivity of structure-guided drug design, but these require computationally expensive molecular dynamics (MD) simulations and rigorous validation [22]. Existing template-based approaches for predicting kinome-wide active-like kinase conformations, although informative, lack benchmarkability because predictions were generated using exclusively active-state templates from the PDB—structures that typically serve as ground truth in evaluation frameworks [4,23]. A more widely adopted strategy for extracting alternative conformations from AlphaFold relies on subsampling multiple sequence alignment (MSA) inputs. Reducing MSA depth has been shown to enhance structural diversity and accuracy in AlphaFold2 models, illustrating the algorithm's capacity to sample multiple conformational states of a given protein [24]. Several heuristic methods have since emerged to improve conformational sampling for alternative conformation prediction (ACP). The AF-cluster approach [25] exploits clustering of MSA sequences by similarity to enable AlphaFold2 to capture alternative states of metamorphic proteins. AFsample2 [26] introduces random masking of MSA columns to attenuate co-evolutionary signals, thereby increasing structural diversity in AlphaFold2-generated ensembles and enabling effective ACP. Guan et al. proposed using the concept of frustration from the energy landscape theory to guide MSA subsampling for predicting conformational structures and motions [27]. Xing et al. developed AF-ClaSeq, which isolates specific co-evolutionary signals through sequence purification by iteratively generating MSA subsets, performing structure predictions, and comparing them to previously sampled conformations to identify novel states [28]. Concurrently, we developed a framework, AFEXplorer (AFEX), that integrates collective variables (CVs)—a core concept in enhanced sampling methodologies—into the challenge of ACP. By reformulating ACP as an optimization problem, AFEX employs a CV-defined loss function to guide conditioned conformational sampling[29].

In this work, we extend this geometry-driven framework to harness AlphaFold2 for predicting active-state conformations of kinases. By integrating critical structural constraints with refined MSA processing, AFEX successfully sampled active-state models for 436 catalytic kinase domains. Benchmarking against recently resolved experimental structures reveals that AFEX achieves substantial improvement over both AlphaFold2 and AlphaFold3 in accurately capturing these functionally crucial conformations. This systematic prediction of kinase activation states opens new avenues for mechanistic insights and has the potential to accelerate structure-based drug discovery across the human kinome.

Results

AFEX Enables Prediction of Active-State Conformations and Ligand Imputation for 436 Human Kinase Domains

Active-state conformations were successfully sampled for 436 catalytic domains across 429 human protein kinases using AFEX optimization, the fundamental geometric constraints guiding the sampling were derived from the structural analysis of all kinase–substrate complexes available in the PDB [4]. The general active-like conformation was defined by the following constraints: a DFG-in conformation, a BLAminus conformation of the XDFG motif, preservation of the salt bridge between the C-helix and β3 strand, and proper extension of the activation loop with contact to the catalytic loop (see Methods). These criteria were validated through our statistical analysis of the

same geometric parameters across all ATP- and ADP-bound kinase structures in the PDB (**Supplementary Figure 1**).

During the AFEX optimization process, a collective loss function was incorporated into the AlphaFold2 inference module, integrating geometric constraints with confidence metrics (pLDDT) for both the global structure and the activation segment (**Figure 1**). Gradient descent was utilized to refine the MSA representation features within AlphaFold2, driving convergence toward conformations that minimized the CV-defined loss. Subsequent statistical analysis of AFEX-predicted structures revealed that both global and activation segment-specific pLDDT scores exceeded 0.95, consistently meeting criteria indicative of the active state (**Supplementary Figure 2**).

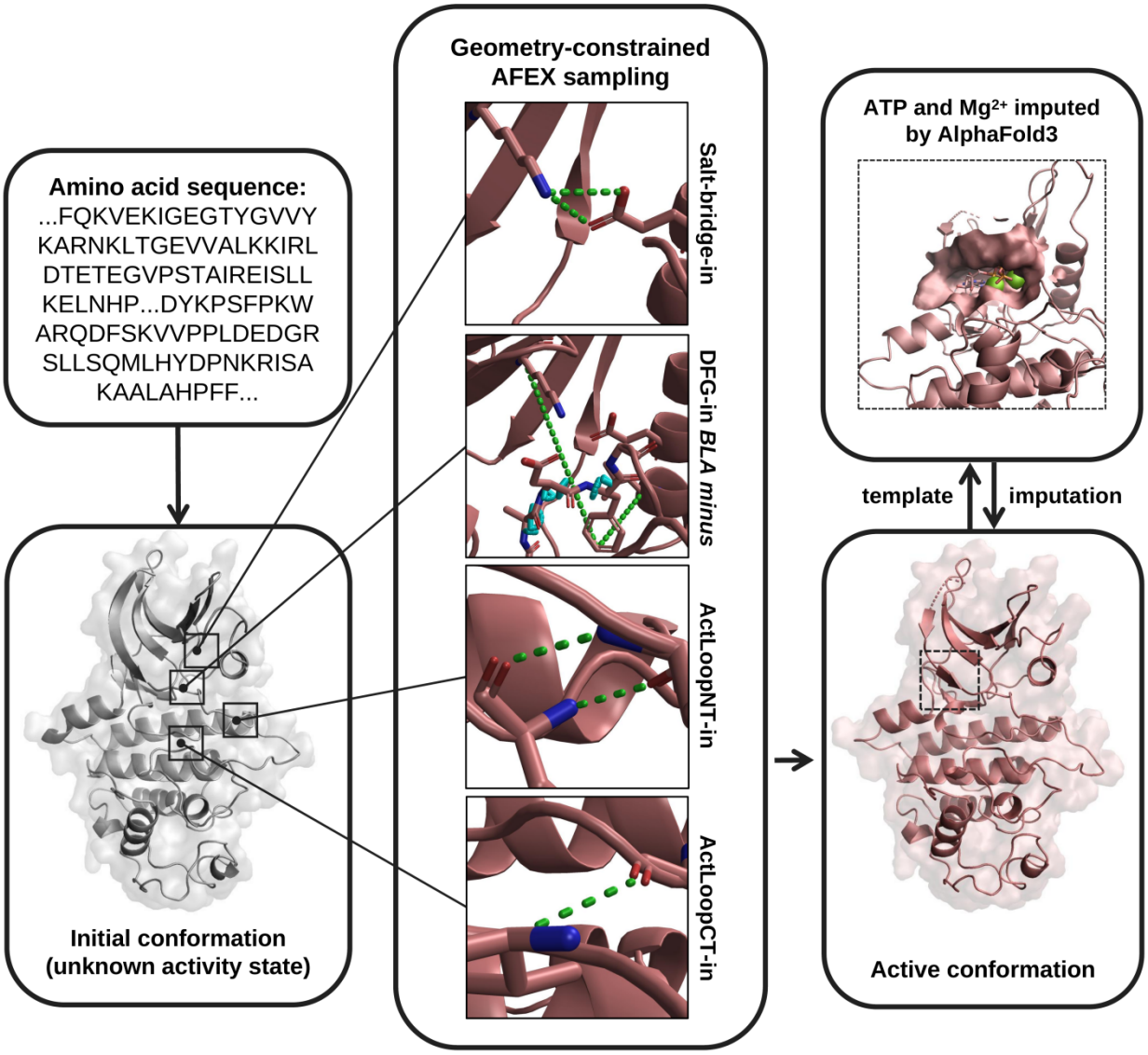

Figure 1. Schematic overview of the AFEXplorer (AFEX) workflow for geometry-constrained prediction of active-state conformations of 436 human kinase domains across the kinome. AFEX systematically samples the conformational space of 436 human kinase domains by iteratively refining MSA representations generated by AlphaFold2. The refinement incorporates explicit geometric constraints that enforce a DFG-in conformation, a BLA-minus conformation of the XDFG motif, preservation of the salt bridge between the C-helix and the β3 strand, and proper extension of the activation loop with contact to the catalytic loop (see Methods). Green dashed lines denote distance constraints imposed between key residues, while cyan dashed lines indicate dihedral angle constraints used to guide critical backbone torsions toward catalytically competent geometries.

We then used the AFEX-predicted active-state structure for each kinase as the sole template in an AlphaFold3 prediction to impute ATP and magnesium ions. The structures were aligned before and after ligand incorporation, and the coordinates of ATP and the two Mg^{2+} ions were transferred from the AlphaFold3 output to the corresponding AFEX structure (see Methods). Minor atomic clashes were resolved through energy minimization in Chimera, yielding AFEX-imputed structures with RMSD values below 0.25 Å relative to the original AFEX predictions (**Supplementary Figure 2e**).

AFEX Outperforms AlphaFold2 and AlphaFold3 in Predicting Active and ATP-Bound Kinase Conformations

To evaluate the accuracy of AFEX predictions, we compared them with 156 experimentally determined substrate-bound active-state kinases curated from 3,010 PDB entries (Roland, biorxiv). AFEX models were generated using the AlphaFold2 framework, trained on PDB structures available before April 2019, and validated against 145 substrate-bound active-state structures released during the same period. AFEX achieved an average RMSD <1 Å for 66.9% of activation segments and 75.9% for global catalytic domains (**Figure 2a and Supplementary Figure 3a**). With the exception of CSF1R, all kinases predicted by AFEX demonstrated RMSDs <2 Å in both their global catalytic domains and activation segments, exhibited higher prediction accuracy for activation segments than Alphafold2 (99.3% vs. 90.3%). Despite being trained on these data, AlphaFold2 failed to generate active-like activation segments for 14 kinases (**Figure 2b**). We note that multiple active-state structures are available for some kinases; therefore, averages and standard deviations are reported for those cases in **Figure 2.**

For 67 kinases with PDB structures of active conformation published after April 2019, AFEX maintained strong performance, achieving RMSD <1 Å in 67.1% of activation segments predictions (**Figure 2c**), with only CDC7 exceeding the 2 Å threshold (**Figure 2d**), a deviation likely due to a long disordered activation loop. AFEX models also exhibited greater consistency across datasets, maintaining approximately 99% accuracy for predictions within 2 Å in both pre- and post-April 2019 data—outperforming AlphaFold2, which achieved 92.5% accuracy and failed to generate active-like conformational states for activation segments of 9 kinases (**Figure 2d**).

To further assess AFEX's utility in modelling ATP-complex structures, we incorporated ATP and two Mg^{2+} binding imputations using AlphaFold3, generating what we refer to as AFEX-impute models (see Methods). These models retain the same backbone conformations as the original AFEX predictions, with ATP and Mg^{2+} positioned through ligand imputation. In AlphaFold3 prediction, ATP and two Mg^{2+} ions were provided in the input, as they are essential for kinase activation. Model performance was assessed relative to AlphaFold3's training data cutoff (September 2021), prior to which 150 active-state conformations are available. AFEX-impute models achieved activation segment predictions with RMSD <1 Å in 65.3% of cases and <2 Å in 98.7% (**and Figure 2e and 2g**). Though AlphaFold3 outperformed AFEX in the <1 Å category (80.7% vs. 65.3%), AFEX maintained superior accuracy in the <2 Å category (98.7% vs. 92.0%) (**Figure 2f**). In this range, AlphaFold3 failed to generate active-like state activation segments for 12 kinases, whereas AFEX failed only for CSF1R and CDC7. More notably, for the subset of 66 kinases with experimentally determined substrate-bound structures published after AlphaFold3's training cutoff, AFEX achieved 71.4% <1 Å accuracy and 100% <2 Å accuracy in activation segment predictions (**Figure 2g**). AlphaFold3 matched AFEX's <1 Å accuracy (71.4% vs. 71.4%) but underperformed at <2 Å (83.3% vs. 100.0%) (**Figure 2h**), failing to generate active-like state activation segments of 7 kinases. These results highlight AFEX's robust performance in modeling ATP-bound kinase conformations following ATP and Mg^{2+} imputation, particularly in maintaining high accuracy in activation segment predictions within the <2 Å threshold.

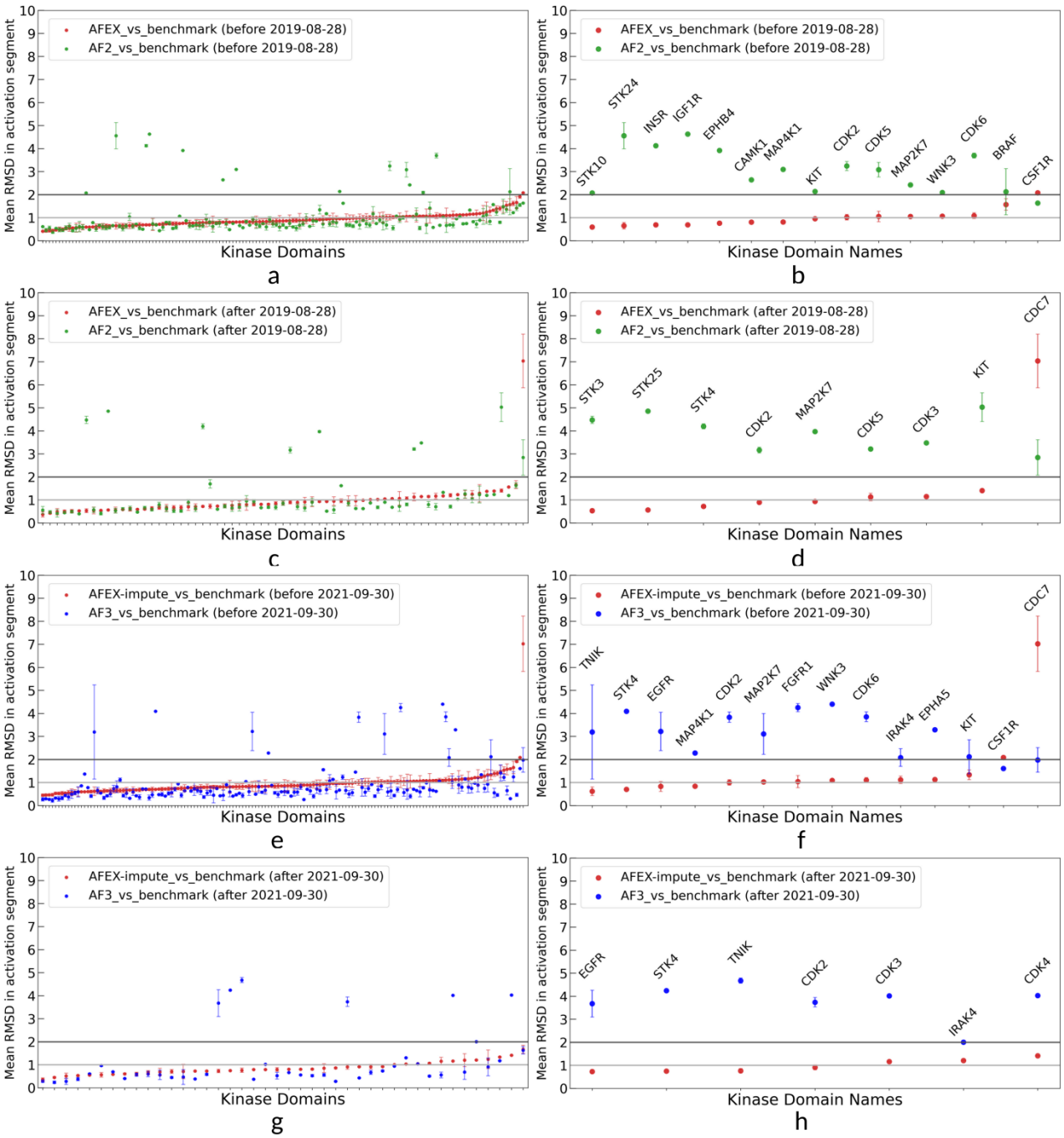

Figure 2. Comparison of activation segments predicted by AFEX and AlphaFold models with experimentally determined active, substrate-bound conformations of 156 different kinases from the Protein Data Bank (PDB). Panels (a)–(d) show the mean root-mean-square deviation (RMSD) and standard deviation (for kinases with multiple active structures in the PDB) of activation segments—considering all RMSD values and those exceeding 2 Å—across residues from the position preceding the DFG motif to the APE motif, evaluated using structures released before and after August 30th 2019, the AlphaFold2 training cutoff date. Panels (e)–(h) present analogous comparisons for structures published before and after September 30th 2021, the AlphaFold3 training cutoff. Results for AFEX and AFEX-impute are shown in red, AlphaFold2 in green, and AlphaFold3 in blue.

Kinases with Extensive Experimental Structural Coverage

Epidermal Growth Factor Receptor (EGFR), a key member of the TYR family and a critical regulator of cellular proliferation, is extensively represented in structural databases, with 267 human EGFR structures deposited in the PDB as cataloged in KLIFS. Of these, 61 active-state conformations were determined from the 128 structures available before the AlphaFold2 training cutoff, and 78 active-state structures were resolved from the 180 structures available prior to the

AlphaFold3 training cutoff. AlphaFold3 predicted inactive-like geometries even when modeled in the presence of ATP and Mg^{2+} (**Figure 3a**), whereas AFEX and AlphaFold2 successfully predicted active-state conformations of EGFR (**Figure 3b**). To elucidate the structural discrepancies among the two AlphaFold models and AFEX-impute, these three structures were superimposed, with a focused analysis on regions exhibiting substantial variation—the activation segment, C-helix, and β3 strand (**Figure 3c**). In the AlphaFold3 models, the activation loop was misoriented toward the ATP-binding pocket, sterically hindering the substrate binding site. As shown in the magnified view, one magnesium ion is slightly displaced from the ATP-binding site in the AlphaFold3-predicted structure (**Figure 3c**). In contrast, when the AFEX-predicted active-like conformation is used as a template, both magnesium ions are accurately coordinated by ATP, reflecting a functionally competent active state.

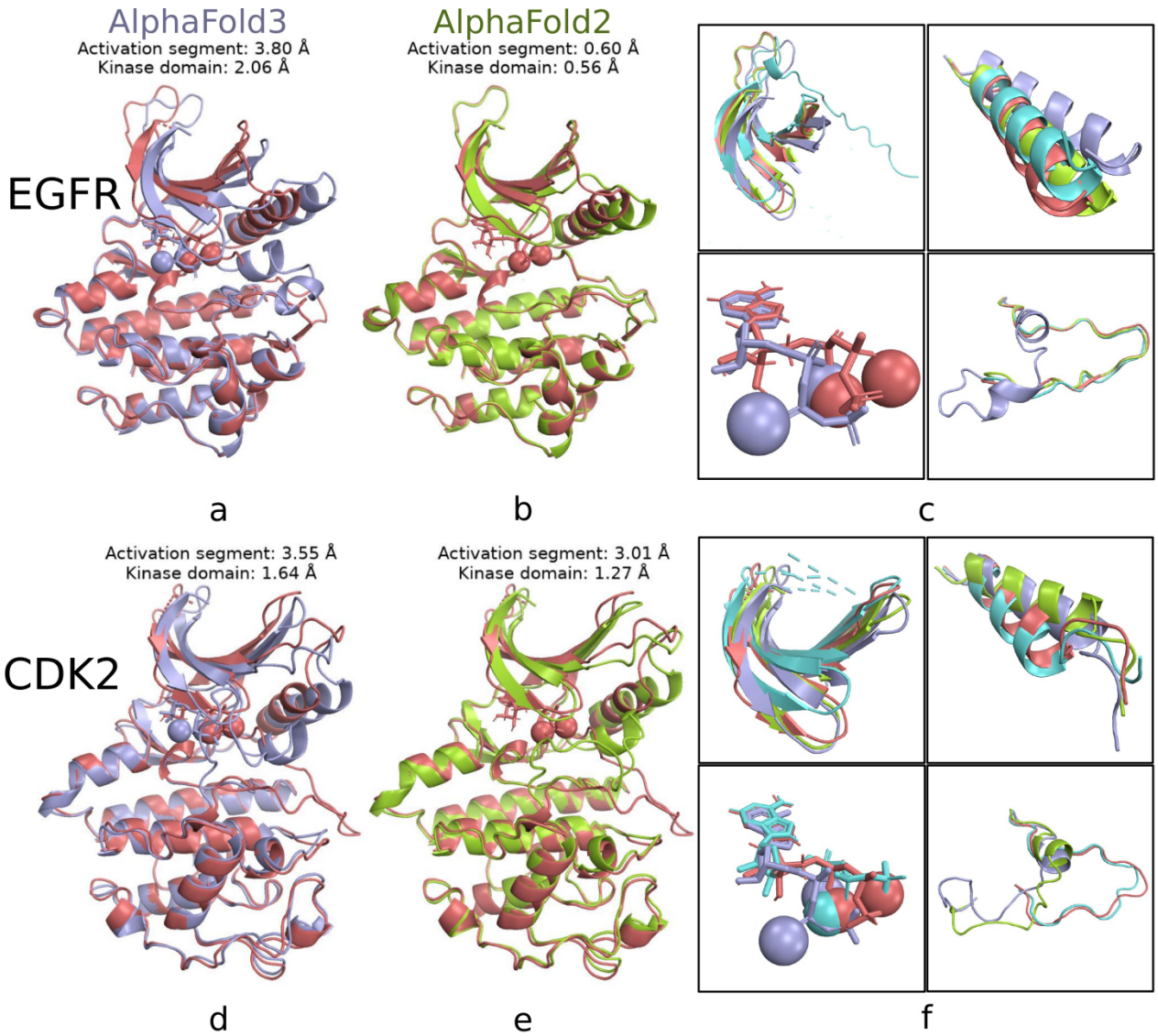

Figure 3. Comparison of EGFR and CDK2 structures predicted by AFEX and AlphaFold models. Panels (a) and (b) show the AFEX-impute prediction of EGFR alongside predictions from AlphaFold3 and AlphaFold2, respectively. Panels (c) displays the superimposed structures, with a detailed view of regions exhibiting significant conformational differences—the β3 strand, the C-helix, bound ligands (ATP and Mg^{2+}), and the activation segment. Panel (d)-(f) present analogous comparisons for CDK2. In all panels, the AFEX-impute predictions are shown in red, the AlphaFold3 models in blue, the AlphaFold2 models in green, and representative substrate-bound PDB structures in aquamarine (EGFR: 2GS2; CDK2: 4EOM).

Cyclin-dependent kinase 2 (CDK2), a key member of the CMGC kinase family involved in cell cycle regulation, is even more extensively represented in the PDB (457 structures). Among these,

110 active-state conformations out of 284 structures were determined prior to the AlphaFold2 training cutoff date, and 115 out of 298 were resolved before the AlphaFold3 training cutoff date. Despite this substantial structural coverage, both AlphaFold2 and AlphaFold3 failed to predict the active state of CDK2, instead modeling the activation loop in substrate-blocking conformations (**Figure 3d-e**). Similar to EGFR, the AlphaFold3 prediction of CDK2 lacks a magnesium ion in the ATP-binding site, even when ATP and two magnesium ions are provided as input, indicating a deficit in functional awareness of AlphaFold3 model (**Figure 3f**). An additional example is BRAF, a representative kinase from the TKL kinase family with 109 structures deposited in the PDB, as illustrated in **Supplementary Figure 4**. Recent studies on fold-switching proteins suggest that AlphaFold models may rely more on memorized structural templates from the training data than on inference grounded in the underlying folding free energy landscape [30]. Under this hypothesis, the presence of too many conflicting conformations in the training set may hinder AlphaFold3's ability to resolve a consistent, functionally relevant active state.

Consistent Prediction of Active-State Conformations in CMGC Kinases

The CMGC family comprises protein kinases involved in several critical signaling pathways, including cell cycle regulation (via cyclin-dependent kinases, CDKs), glycogen metabolism (via glycogen synthase kinase 3, GSK3), stress and growth responses (via mitogen-activated protein kinases, MAPKs), and RNA processing (via SR protein kinases). These kinases primarily catalyze the phosphorylation of serine and threonine residues and play pivotal roles in cellular proliferation, differentiation, and apoptosis, with select members also displaying tyrosine kinase activity. When we superimposed all the predicted CMGC structures together, the AFEX model - unlike the diverse and sometimes closed conformations of the activation segments predicted by AlphaFold2 and AlphaFold3 (**Figure 4a and 4b**) - revealed a distinctly open conformation of the ATP and substrate binding site (**Figure 4c**). This feature was also consistently observed across other serine/threonine kinase (STK) families upon structural superposition (**Supplementary Figure 5**).

In our benchmark analysis, all CMGC kinases for which AlphaFold failed to achieve 2 Å accuracy belonged to the CDK subfamily (**Figure 2**). Both AlphaFold2 and AlphaFold3 predicted inactive conformations for CDK2, CDK3, and CDK6, despite the availability of experimentally determined active-state structures for these kinases during training. For CDK5, although the six experimentally determined structures in the PDB prior to the training cutoff all represented the active state, AlphaFold2 predicted an inactive conformation characterized by a folded activation loop, disruption of the salt bridge between the C-helix and β3 strand, and a DFG motif in the BLB-minus conformation (**Figure 4e–f**). In contrast, AlphaFold3 predicted an active-state conformation for CDK5 (**Figure 4d**), suggesting that the inclusion of ATP and Mg^{2+} in the modeling process enhanced the prediction of active kinase states—a finding also observed for BRAF(**Supplementary Figure 4d-f**) and CAMK1 (**Supplementary Figure 6a–c**), a calcium/calmodulin-dependent serine/threonine kinase from the CAMK family. Similar improvements were evident in the STK kinases from the STE family: AlphaFold2 predicted inactive conformations for STK3, STK10, STK24, and STK25, whereas AlphaFold3 successfully predicted active-state conformations when ATP and Mg^{2+} ions were included (**Supplementary Figure 7**).

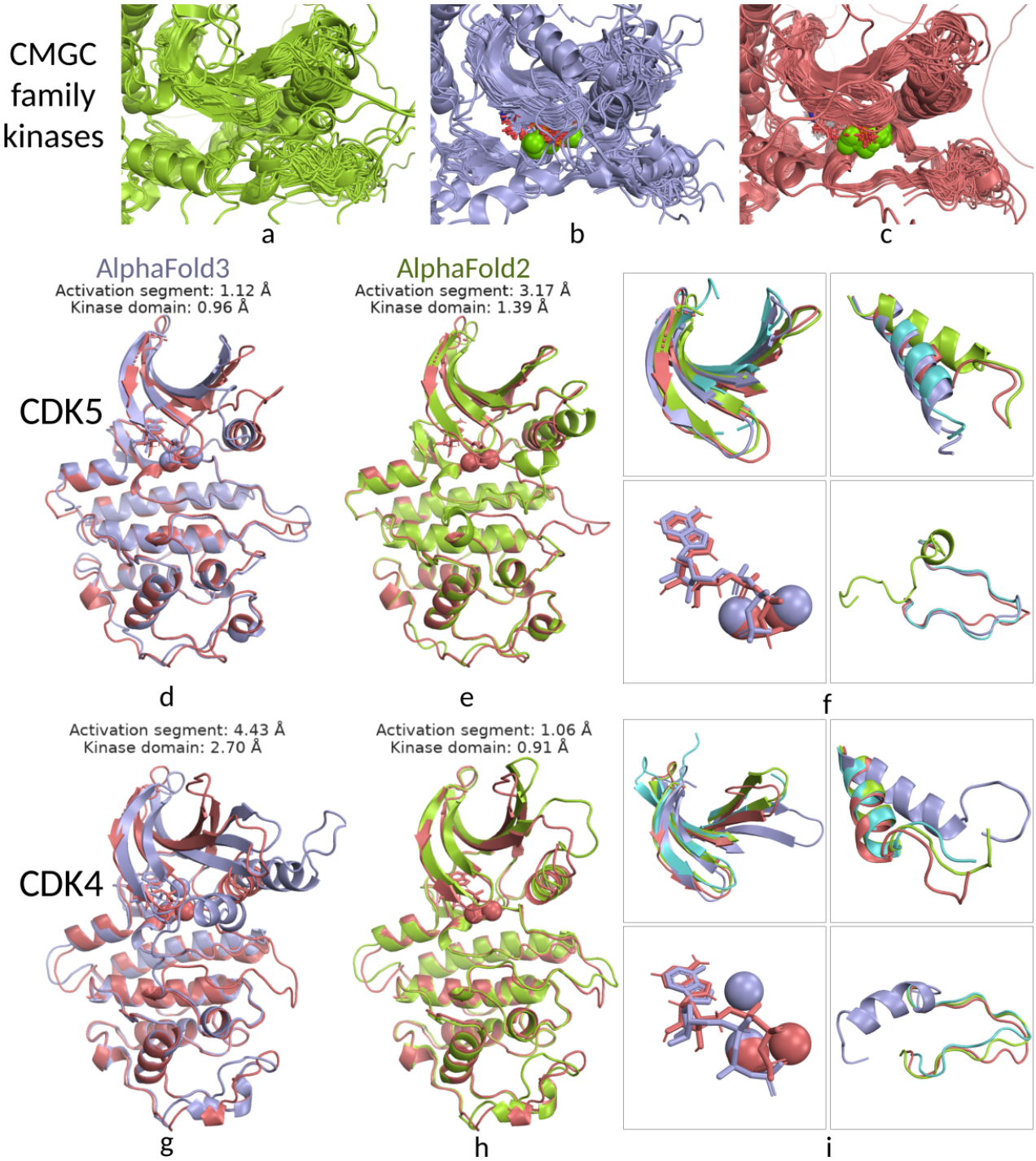

Figure 4. Comparison of CMGC family kinases, CDK5, and CDK4 structures predicted by AFEX and AlphaFold models. Panels (a), (b) and (c) display the superimposed structures of CMGC family kinases predicted by Alphafold2, AlphaFold3, and AFEX-impute, respectively, with focus on regions exhibiting significant conformational differences. Panels (d)-(e) show the AFEX-impute prediction of CDK5 alongside predictions from AlphaFold3 and AlphaFold2, respectively. Panels (f) presents the superimposed CDK5 structures with detailed focus on the β3 strand, the C-helix, bound ligands (ATP and Mg^{2+}), and the activation segment, respectively. Panel (g)-(i) present analogous comparisons for CDK4. In all panels, the AFEX-impute predictions are shown in red, the AlphaFold3 models in blue, the AlphaFold2 models in green, and representative substrate-bound PDB structures in aquamarine (CDK5: 1UNG; CDK4: 7SJ3).

However, AlphaFold3 failed to recover the active state of CDK4 in the presence of ATP and Mg^{2+} (**Figure 4g**), mispositioning the activation loop and failing to form the essential salt bridge between the C-helix and the β3 strand—a finding also observed for STK4 (**Supplementary Figure 8a–c**). In the case of CDK4, although neither model was trained on active-state structures of CDK4,

AlphaFold2 successfully predicted the active conformation (**Figure 4h**), while AlphaFold3 did not (**Figure 4g and 4i**). This discrepancy highlights the stochastic nature of active-like conformation prediction when including ATP and Mg^{2+} in the input; the emergence of an open binding site is not guaranteed. A similar pattern was observed in EGFR (**Figure 3c**) and TNIK (**Supplementary Figure 8d–f**), a member of the STE kinase family, where the inclusion of ATP and Mg^{2+} resulted in a closed substrate-binding site. Interestingly, WNK3, an orphan kinase, exhibited an intermediate behavior: AlphaFold2 predicted a helical conformation in the activation loop, whereas AlphaFold3, when supplied with ATP and two magnesium ions, produced an unwound helix but lost the interaction between the N-terminal region of the activation loop and the catalytic loop, ultimately failing to predict the active state (**Supplementary Figure 9a-c**). In contrast, we highlight that AFEX-impute consistently predicted active-state conformations and generated structurally plausible coordinates for ATP and both Mg^{2+} ions across both CDKs (**Figure 4f and 4i**) and STKs (**Supplementary Figure 7 and 8c**), achieving RMSD values below 2 Å in available benchmarks (18 kinases, **Figure 2**).

AFEX-Optimized Templates Facilitate Magnesium Ion Placement in STE Family Kinase Models

The STE kinase family was initially identified in the MAPK signaling pathway in yeast and was later shown to serve as upstream regulators in mammalian MAPK pathways. Notably, AlphaFold3 encounters difficulties in accurately modeling ATP and magnesium ion binding within the STE kinase family compared to other kinase families. Specifically, in five of the 45 catalytic domains—MAP2K3, MAP2K4, MAP2K6, MAP2K7, and TNIK—the magnesium ion is misplaced outside the ATP-binding site (**Supplementary Figure 10a**). For MAP2K4 (**Figure 5a–c**) and MAP2K6 (**Supplementary Figure 8g–i**), structurally plausible positioning of the magnesium ions was achieved only when AFEX optimized active-like structures were used as templates. This inaccuracy of ion placement in the AlphaFold models is likely due to the formation of an excessively helical activation loop, which may sterically hinder the proper coordination of the magnesium ions. A similar improvement was observed in WNK1, where the use of AFEX-predicted active-state models enabled AlphaFold3 to correctly open the ATP-binding site and position the magnesium ions accurately (**Supplementary Figure 9d-f**).

Both AlphaFold2 and AlphaFold3 predicted inactive conformations of MAP2K7 (MKK7), characterized by a helical segment within the activation loop, the absence of a salt bridge between the C-helix and the β3 strand, and an inactive DFG motif conformation (**Figure 5d-f**). Although AFEX successfully predicted the active-state structure in agreement with the experimentally determined substrate-binding conformation, AlphaFold3 failed to accurately predict the positions of magnesium ions in MAP2K7 using the AFEX-predicted active-like structure as a template. Similar failures were observed in MAP2K3 (**Supplementary Figure 8j–l**) and TNIK (**Supplementary Figure 8d-f**), indicating persistent challenges of AlphaFold3 in predicting magnesium ion coordination within the STE kinase family.

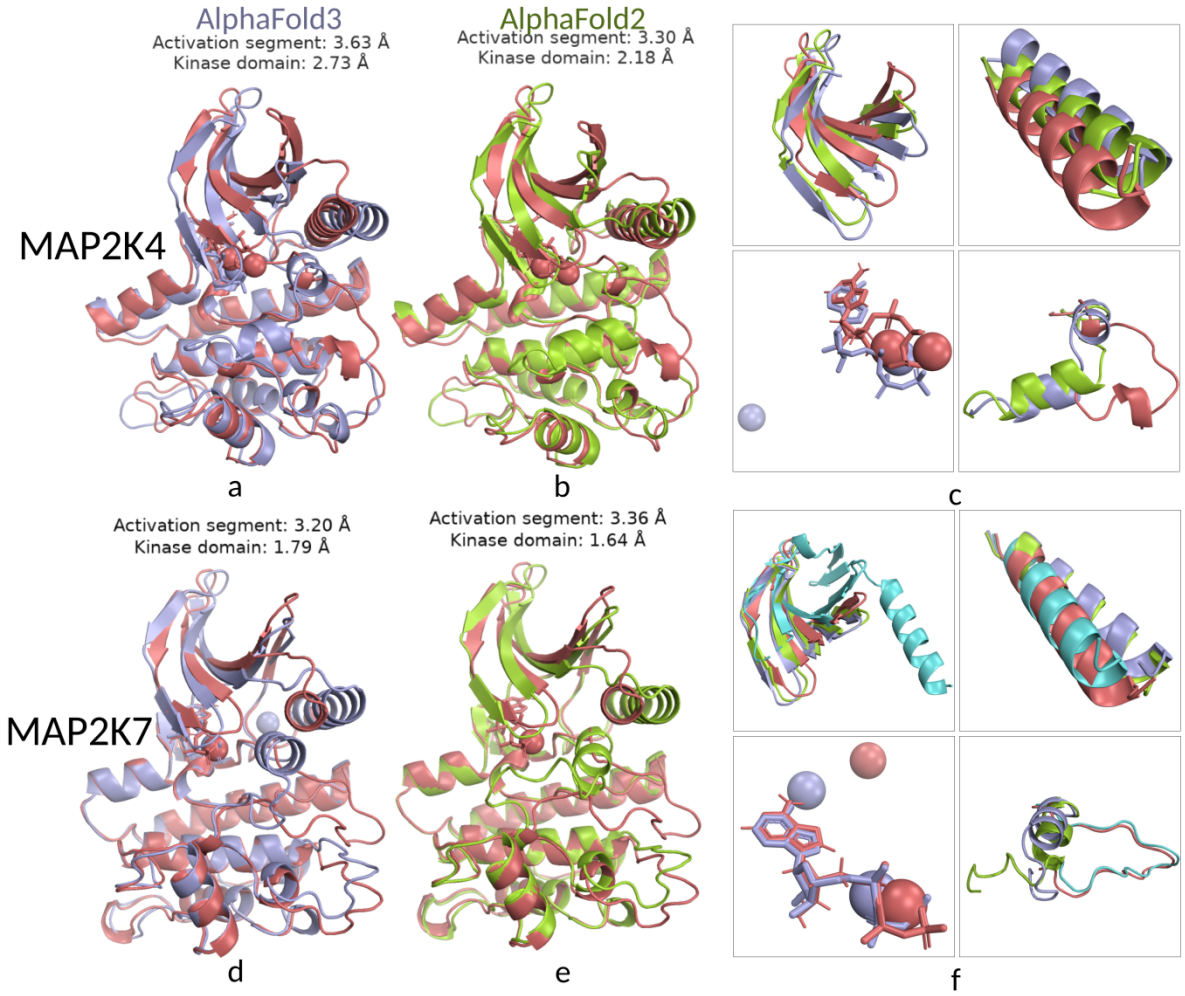

Figure 5. Comparison of MAP2K4 and MAP2K7 structures predicted by AFEX and AlphaFold models. Panels (a) and (b) show the AFEX-impute prediction of MAP2K4 alongside predictions from AlphaFold3 and AlphaFold2, respectively. Panels (c) displays the superimposed structures of MAP2K4, with a detailed focus on regions exhibiting significant conformational differences—the β3 strand, the C-helix, bound ligands (ATP and Mg^{2+}), and the activation segment. Panel (d)-(f) present analogous comparisons for MAP2K7. In all panels, the AFEX-impute predictions are shown in red, the AlphaFold3 models in blue, the AlphaFold2 models in green, and representative substrate-bound PDB structures in aquamarine (MAP2K7: 6YG1).

AFEX Consistently Predicts Open Binding Site Conformations for TYR Kinases and Rescues ATP Imputation in AATK

The TYR family comprises classical tyrosine protein kinases that catalyze the phosphorylation of tyrosine residues in proteins. This family includes both receptor tyrosine kinases, such as EGFR and INSR, and intracellular tyrosine kinases, such as SRC. The 82 catalytic kinase domains of all TYR family members were structurally aligned and compared (**Supplementary Figure 11**). Similar to observations in serine and threonine kinases, AlphaFold2 and AlphaFold3 predicted multiple closed conformations of the activation segments, whereas the AFEX structures revealed a distinct open conformation of the ATP and substrate binding site.

While some TYR kinases such as EGFR have extensive experimental structural coverage, others in this family, for example AATK, lack direct structural data in the PDB. AlphaFold2 predicts that the activation loop of AATK adopts a helical, inactive conformation (**Figures 6a**). In AlphaFold3, even with the inclusion of two magnesium ions and ATP as input, the model failed to adopt an active-like conformation; instead, ATP was mispositioned between the activation loop and the C-helix (**Figure 6a**). Only when the AFEX-predicted active-state structure was supplied as a template did

AlphaFold3 successfully generate a model with plausible binding modes for both ATP and the magnesium ions (**Figure 6c**).

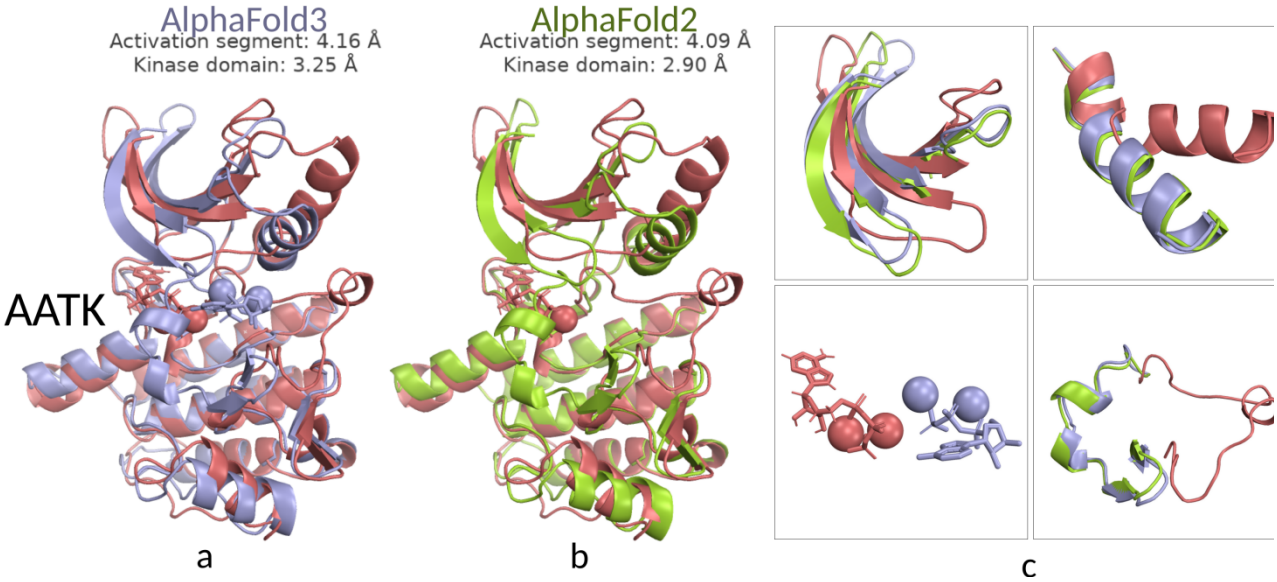

Figure 6. Comparison of TYR family structures and AATK structures predicted by AFEX and AlphaFold models. Panels (a) and (b) show the AFEX-impute prediction of AATK alongside predictions from AlphaFold3 and AlphaFold2, respectively. Panels (c) displays the superimposed structures of AATK, with a detailed focus on regions exhibiting significant conformational differences—the β3 strand, the C-helix, bound ligands (ATP and Mg^{2+}), and the activation segment. In all panels, the AFEX-impute predictions are shown in red, the AlphaFold3 models in blue, the AlphaFold2 models in green.

Conclusion and Discussion

In conclusion, our results demonstrate that AFEX reliably predicts active-state kinase conformations across diverse kinase families in a user-controllable manner, achieving structural accuracy sufficient to accommodate ATP and two magnesium ions within catalytic binding sites in most cases—outperforming predictions generated by AlphaFold3 alone. AFEX offers a solution to a critical limitation of current methodologies: while AlphaFold2 and AlphaFold3 often fail to generate active-state conformations, AFEX consistently produces functionally relevant geometries. The robustness of AFEX is particularly evident in its accurate modeling of activation loops and substrate-binding regions, which are sometimes mispositioned in AlphaFold predictions, leading to steric occlusion of substrates. By preserving the functional architecture of the ATP-binding site, AFEX provides a kinome-wide framework for investigating kinase activation mechanisms and supporting structure-based design of kinase inhibitors.

This work also addresses the broader limitations of current structure prediction models in capturing conformationally dynamic protein states. Benchmark analysis revealed that AlphaFold2 and AlphaFold3 failed to predict catalytically competent conformations in a substantial number of kinases, including eight from the TYR family, eight from the STE family, five from the CMGC family, two from the TKL family, and one each from the CAMK and Other families, with RMSD values exceeding 2 Å (**Figure 2**). These mispredictions occurred even for kinases with abundant experimental active-state structural data in the PDB (**Figure 3**). Recent studies on fold-switching proteins suggest that such failures might stem from AlphaFold's reliance on memorized structural templates rather than generalizable principles derived from the protein folding free energy landscape [30]. Notably, AlphaFold's accuracy for known fold-switching proteins remains below 35%, with high false-negative rates regardless of training set inclusion [30,31]. Similar conclusions were drawn by Schafer et al., who showed that AlphaFold2's improved performance over a reduced-training model stems from access to a broader repertoire of memorized templates rather than superior modeling of biophysical principles; both models displayed limited capacity to sample

alternative conformations [32]. Together, these findings highlight the potential advantage for geometry-constrained approaches like AFEX, which reduce interpolation bias and enforce biophysically meaningful constraints to improve predictive accuracy for functional states.

Our findings also show that prediction quality does not correlate with the availability of active-state training data. Although AlphaFold2 predominantly predicts active-state kinases, with approximately 70% adopting the DFG-in conformation [23,33], only 48% of human catalytic kinases in the EBI AlphaFold dataset exhibit catalytically competent conformations [4]. For well-characterized kinases such as EGFR and CDK2, AlphaFold3 generated inactive-state conformations despite the abundance of structural data. The deviation of AlphaFold predictions from the geometrically defined active state in certain serine/threonine-specific kinases is unlikely to result from insufficient training data, as serine and threonine kinases possess substantially more experimentally determined active-state structures compared to tyrosine kinases [2,34]. This observation challenges assumptions that training data abundance ensures correct functional predictions and highlights the importance of structural constraints in guiding conformational sampling. By incorporating features such as activation loop extension, DFG-in geometry, and C-helix-β3 salt bridge formation, AFEX produces diverse, family-specific active-state models that align well with experimental distributions while allowing flexibility for kinase-specific variations (**Supplementary Figure 2**). It accurately distinguishes the extended activation loops of TYR family kinases from the compact conformations of serine/threonine kinases and adapts to C-helix-β3 salt bridge variations in TKL kinases. Furthermore, AFEX accommodates atypical or orphan kinases by adapting its criteria accordingly, such as omitting salt bridge evaluation for WNK kinases lacking a β3 lysine, enabling accurate modeling across the structural diversity of the kinome.

The availability of active-state models generated by AFEX opens multiple avenues for further computational investigations. They can serve as starting points for MD simulations to study the conformational dynamics of kinase and how these dynamics are modulated by by interacting ligands. The resolution of both inactive and active states is particularly useful for path-based methods, such as constant advance replicas [35,36] and finite-temperature string [37] methods. AFEX-impute provides accurate structures of ATP- and Mg^{2+}-bound kinase complexes across the human catalytic kinome, which could serve as suitable input structures for QM/MM simulations to probe catalytic mechanisms and chemical reactivity in a biologically relevant state. This enables mechanistic studies of activation dynamics, and facilitates kinetic modeling of signaling processes. Beyond individual kinases, these predictions may support structural inference of higher-order assemblies and complexes involving kinase domains—particularly where conformational plasticity underlies functional regulation.

Furthermore, the availability of functionally accurate active-state models may also facilitate the identification of kinase-substrate pairs, which is essential for elucidating cellular signaling pathways but remains a significant challenge—particularly for "dark" kinases that lack functional annotation [38]. Traditional biochemical methods, such as affinity chromatography coupled with mass spectrometry, have been instrumental in mapping phosphorylation sites and uncovering kinase-substrate complexes [39,40]. Yet these approaches have left over 90% of known phosphosites unassigned to a specific kinase, and nearly 30% of human kinases without any experimentally validated substrate [38], highlighting the scale of the knowledge gap and the need for scalable and accurate computational solutions.

While sequence-based computational methods have advanced kinase-substrate predictions [41,42], their performance relies heavily on curated databases such as PhosphoSitePlus [43] and Phospho.ELM [44], which are incomplete and skewed toward well-characterized kinases, thereby limiting generalizability and predictive accuracy [45]. More importantly, these models often overlook the structural elements that determine substrate recognition. Structure-informed approaches, such as

SAGEPhos [46], have started to incorporate 3D features to improve prediction quality. However, most available structural data lack catalytically competent kinase conformations: only a small fraction of kinase structures in public databases adopt active-like states [4]. Integrating functional conformational states into prediction pipelines, enabled by the current work, offers a promising strategy to address these limitations. In general, by anchoring prediction models in functionally relevant structural contexts, AFEX has the potential to enhance kinase–substrate mapping and support discovery of previously uncharacterized components of signaling networks.

In summary, AFEX bridges a critical gap in current biomolecular structural prediction by enabling functionally accurate, conformation-aware representations of kinase active states. It complements existing deep learning models and provides a generalizable framework applicable to kinases across species and other protein families where functional states are defined by conserved activation signatures, including GPCRs and ion channels. To support community use and further analysis, all AFEX-predicted active-state models for the human kinome have been made publicly available. While further refinement may be needed to address the dynamics of disordered loops, the integration of AFEX with other ML-based or simulation-based approaches holds promise for expanding the mechanistic toolkit of kinase biology and structure-based drug discovery.

Methods

Geometry-constrained Model Generation with AFEXplorer

AFEX samples kinase conformations by defining active-state prediction as a CV-driven problem to bias the generation of structures with AlphaFold2. In particular, AFEX employed five geometric criteria to characterize active kinase conformations, adapted from structural analyses of kinase–substrate complexes [4]:

Active Position of the DFG Motif (DFG-in)

The phenylalanine (Phe) residue of the DFG motif occupies the "DFG-in" position beneath the C-helix or within a proximal pocket in the N-terminal lobe. This is quantified by two distance constraints: the Cα atom of the fourth residue following the conserved glutamate (Glu) on the C-helix (i.e., the residue that is four positions C-terminal to the Glu, denoted as Glu4-Cα) to the Cζ atom of DFG Phe (Phe-Cζ) must be less than 11 Å ($d_1 < 11$ Å), and the Cα atom of the conserved β3 lysine (Lys-Cα) to Phe-Cζ must range between 11 and 20 Å (11 Å $\leq d_2 \leq$ 20 Å) [47].

$$L_{d_1} = \max(0, 0 - d_1) + \max(0, d_1 - 11) \quad (1)$$
$$L_{d_2} = \max(0, 11 - d_2) + \max(0, d_2 - 20) \quad (2)$$

XDFG Dihedral Angles and BLAminus Conformation

The χ_1 dihedral angle of the N, Cα, Cβ, Cγ of the Phe residue of DFG motif adopts a gauche- (g^-) rotamer, corresponding to values between 240° and 360°:

$$L_{\chi_1} = (\max(0, 240^\circ - \chi_1) + \max(0, \chi_1 - 360^\circ)) \times \frac{\pi}{180^\circ}, \ (\chi_1 \in [0^\circ, 360^\circ]) \quad (3)$$

Additionally, the backbone dihedral angles (ϕ and ψ) of the one preceding the DFG motif ("X"), Asp, and Phe residues in the XDFG motif must align with the "BLAminus" region of the Ramachandran plot [2]. This is assessed using cosine distance, defined for each cluster *k* in {BLAminus, ABAminus, BLBminus} as:

$$d_{cos}^{(k)} = \frac{2}{6}\sum_{j=1}^{6}\left[1 - \cos(\alpha_j^{pred} - \alpha_j^{(k)})\right] \quad (4)$$

where j represents the six dihedral angles (ϕ and ψ) of the "X", Asp, and Phe residues in the XDFG motif, respectively. α_j^{pred} denotes the predicted value, and $\alpha_j^{(k)}$ represents the corresponding predefined ideal angle for cluster k [4]. The loss $L_{XDFG}^{BLAminus}$ for the 'BLAminus' cluster is then formulated as:

$$L_{XDFG}^{BLAminus} = \sum_{k=2}^{n} max\left(0, d_{cos}^{(1)} - d_{cos}^{(k)}\right) + max\left(0, d_{cos}^{(1)} - 0.45\right) \quad (5)$$

C-helix-β3 Salt Bridge Distance (Salt-bridge-in)

In the active state, a salt bridge forms between the conserved glutamate (or substitute) on the C-helix and the lysine (or substitute) in the β3 strand, stabilizing the ATP-binding site. The salt-bridge distance (d_{sb}) is defined as the minimum distance between any hydrogen-bonding atoms in these residues, typically the Lys-N and Glu-O atomic distances must be less than 3.6 Å. For members of the TKL family, this threshold is relaxed to 5.1 Å. No constraint was applied when the β3 lysine is substituted to asparagine (Asn), as this substitution disrupts the salt bridge:

$$L_{sb} = \begin{cases} 0, \text{Asn substitution of Lys} \\ max(0, d_{sb} - 5.1), \text{TKL family kinases} \\ max(0, d_{sb} - 3.6), \text{other non-TKL family kinases} \end{cases} \quad (6)$$

Activation Loop C-terminal Position (ActLoopCT-in)

The C-terminal segment of the activation loop engages the substrate in the active state. Thus, the distance between the Cα atom of the residue nine positions upstream of the APE motif (APE9-Cα) and the backbone carbonyl oxygen of the arginine in the HRD motif (Arg-O) must be less than 6.0 Å for non-tyrosine kinase (non-TYR) families and less than 8.0 Å for TYR family kinases [4]:

$$L_{CT} = \begin{cases} max(0, d_{CT} - 6), \text{non-TYR family kinases} \\ max(0, d_{CT} - 8), \text{TYR family kinases} \end{cases} \quad (7)$$

Activation Loop N-terminal Position (ActLoopNT-in)

The N-terminal region of the activation loop forms a backbone hydrogen bond critical for catalytic competence. This interaction occurs between the sixth residue of the DFGxxX motif (DFG6) and the residue immediately preceding the HRD motif (XHRD). The hydrogen bond is formed between the backbone amide nitrogen (N) of DFG6 and the backbone carbonyl oxygen (O) of XHRD, or *vice versa*. The minimal distance between donor and acceptor atoms (DFG6-N/O to XHRD-N/O) must be less than 3.6 Å:

$$d_{NT} = min(\|N_{XHRD} - O_{DFG6}\|, \|O_{XHRD} - N_{DFG6}\|) \quad (8)$$

$$L_{NT} = max(0, d_{NT} - 3.6) \quad (9)$$

To ensure model reliability, a pLDDT-based loss term was applied to both the catalytic domain and the activation loop segment (spanning the DFG to APE motifs). The loss is zero for per-residue pLDDT scores mean≥0.95 and increases linearly for lower confidence values:

$$L_{pLDDT} = max(0, 0.95 - pLDDT) \quad (10)$$

The total loss function for AFEX optimization is the sum of all geometric and confidence-based terms. These geometric criteria are added together to form the following loss function for AFEX optimization:

$$L_{total} = L_{d_1} + L_{d_2} + L_{sb} + L_{CT} + L_{NT} + 10 \times \left(L_{\chi_1} + L_{XDFG}^{BLAminus}\right) + 2 \times L_{pLDDT} + L_{pLDDT_{actloop}} \quad (11)$$

Models were deemed optimized when all five geometric criteria were satisfied and the average pLDDT score exceeded 0.95 (i.e., $L_{total} = 0$). Although 50 refinement iterations were executed, final models were selected from the 25th iteration onward to ensure convergence, with RMSD between the 25th and 50th iteration structures consistently below 0.1 Å. We also note that the following exceptions to geometric criteria were considered for certain kinases due to their unique structural features:

APE Motif Variants

Kinases such as BUB1, HASPIN, TP53RK, and PKDCC lacked a canonical APE motif, resulting in altered C-terminal activation loop architecture. These were excluded from ActLoopCT-in requirements.

Long Disordered Activation Loop

MASTL kinase was excluded due to an exceptionally long disordered activation loop (575 amino acids), which fell outside the study's scope and was therefore excluded from analysis.

Kinase Structure Prediction Using AlphaFold

The structures of human protein kinase catalytic domains were predicted using AlphaFold2 [15] and AlphaFold3 [16], developed by Google DeepMind. For AlphaFold2 predictions, the monomer model was used. The FASTA formatted amino acid sequences for kinase catalytic domains were obtained from the Kincore database [4] and used for input. When predictions were conducted in template-disabled mode in AlphaFold2, the "--max_template_date" was set to 1700-01-01 with a reduced database preset (--db_preset reduced_dbs). Precomputed MSA function was enabled (--use_precomputed_msas), and all other parameters were kept at their default values. For AlphaFold3 predictions, input amino acid sequences were provided in JSON format.

Imputation of ATP and Mg^{2+} ions

To impute ATP and Mg^{2+} ions into the AFEX models, the bound ligands were incorporated into the input JSON file, and AlphaFold3 was configured to use the AFEX model as the sole structural template. Ligands and ions were subsequently imputed into the active-state kinase domains by aligning the AlphaFold3-predicted complexes with the AFEX models for *in situ* imputation. The kinase structures containing ligands and ions underwent default structural energy minimisation using Chimera 1.18 [48] to potential eliminate atomic clashes. Key residues critical for active-state conformations were frozen during minimisation to preserve structural integrity. These residues included the XHRD and XDFG motifs, Glu and Glu4 on the C-helix, β3 lysine, APE9, and DFG6. After energy minimisation, both the pre- and post-imputed AFEX kinase structures were confirmed to retain active-state conformations using Kincore [23].

Benchmark test

The predicted active-state models were validated by comparing them against 156 PDB structures that met substrate-binding activity criteria. Validation relied on root-mean-square deviation (RMSD) calculations to measure the average atomic displacement between predicted and experimentally resolved structures. The alignment and RMSD computation were conducted using PyMOL's automated multi-step superposition algorithm. This algorithm employs sequence alignment based on BLOSUM62 weightings [49] and only one cycle of iterative refinement to reject outliers, focusing on backbone atoms (N, Cα, C, O). After aligning the model and reference structures, RMSD values for the activation segment backbone atoms were calculated using the "rms_cur" function in PyMOL, which computes RMSD without performing additional fitting.

Data Availability and Reproducibility

The data and source code that support the findings of this study are available at https://github.com/JingHuangLab/AFEX-kinome. The FASTA files of sequences used to model the 436 human catalytic protein kinase domains are available through the Kincore database at https://dunbrack.fccc.edu/kincore/static/downloads/af2activemodels/kinasedomainfasta.tar.gz. The benchmarking dataset comprising 3,010 experimentally determined active structures of 156 human catalytic protein kinases from the PDB used in this study is provided at https://github.com/JingHuangLab/AFEX-kinome/releases/tag/benchmark-2025-10-23. The AFEX active-state structural models for all 436 catalytic kinases, including those with imputed ATP- and magnesium-bound states, are publicly accessible at https://github.com/JingHuangLab/AFEX-kinome/tree/main/database.

ACKNOWLEDGMENTS

We thank Dr. Zilin Song, Dr. Tengyu Xie, Dr. Jing Xue, and Ms. Runtong Qian for insightful discussions. This work is supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2023C03109, 2024SSYS0036), the National Natural Science Foundation of China (32171247), the Westlake University-Muyuan Joint Research Institute (WU2025MY001), the State Key Laboratory of Gene Expression, and the Westlake Education Foundation. We thank the Westlake University Supercomputer Center for computational resources and related assistance.

References

1. Alganem, K. et al. The active kinome: The modern view of how active protein kinase networks fit in biological research. Curr Opin Pharmacol 62, 117–129 (2022).
2. Modi, V. & Dunbrack, R. L. Defining a new nomenclature for the structures of active and inactive kinases. Proc Natl Acad Sci U S A 116, 6818–6827 (2019).
3. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. Science 298, 1912–1934 (2002).
4. Faezov, B. & Dunbrack, R. L. AlphaFold2 models of the active form of all 437 catalytically competent human protein kinase domains. bioRxiv 2023.07.21.550125 (2023) doi:10.1101/2023.07.21.550125.
5. Taylor, S. S., Shaw, A. S., Kannan, N. & Kornev, A. P. Integration of signaling in the kinome: Architecture and regulation of the αC Helix. Biochim Biophys Acta 1854, 1567–1574 (2015).
6. Möbitz, H. The ABC of protein kinase conformations. Biochim Biophys Acta 1854, 1555–1566 (2015).
7. Kugler, V. et al. Impact of protein and small molecule interactions on kinase conformations. Elife 13, RP94755 (2024).
8. Gough, N. R. & Kalodimos, C. G. Exploring the conformational landscape of protein kinases. Curr Opin Struct Biol 88, 102890 (2024).
9. Liljedahl, M. et al. Protein kinase D regulates the fission of cell surface destined transport carriers from the trans-Golgi network. Cell 104, 409–420 (2001).
10. Baron, C. L. & Malhotra, V. Role of diacylglycerol in PKD recruitment to the TGN and protein transport to the plasma membrane. Science 295, 325–328 (2002).
11. Lidke, D. S. et al. ERK nuclear translocation is dimerization-independent but controlled by the rate of phosphorylation. J Biol Chem 285, 3092–3102 (2010).
12. Eitel, K. et al. Protein kinase C delta activation and translocation to the nucleus are required for fatty acid-induced apoptosis of insulin-secreting cells. Diabetes 52, 991–997 (2003).
13. Nakano, R., Kitanaka, T., Namba, S., Kitanaka, N. & Sugiya, H. Protein kinase Cε regulates nuclear translocation of extracellular signal-regulated kinase, which contributes to bradykinin-induced cyclooxygenase-2 expression. Sci Rep 8, 8535 (2018).
14. Greber, B. J. High-resolution cryo-EM of a small protein complex: The structure of the human CDK-activating kinase. Structure S0969-2126(24)00085–6 (2024) doi:10.1016/j.str.2024.03.003.
15. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).
16. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630, 493–500 (2024).
17. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444 (2022).
18. Lyu, J. et al. AlphaFold2 structures guide prospective ligand discovery. Science 384, eadn6354 (2024).
19. Díaz-Rovira, A. M. et al. Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures. J Chem Inf Model 63, 1668–1674 (2023).
20. Scardino, V., Di Filippo, J. I. & Cavasotto, C. N. How good are AlphaFold models for docking-based virtual screening? iScience 26, 105920 (2023).
21. Holcomb, M., Chang, Y.-T., Goodsell, D. S. & Forli, S. Evaluation of AlphaFold2 structures as docking targets. Protein Sci 32, e4530 (2023).
22. Gu, X., Aranganathan, A. & Tiwary, P. Empowering AlphaFold2 for protein conformation selective drug discovery with AlphaFold2-RAVE. Elife 13, RP99702 (2024).
23. Modi, V. & Dunbrack, R. L. Kincore: a web resource for structural classification of protein kinases and their inhibitors. Nucleic Acids Res 50, D654–D664 (2022).

24. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with AlphaFold2. Elife 11, e75751 (2022).
25. Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. Nature 625, 832–839 (2024).
26. Kalakoti, Y. & Wallner, B. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. Commun Biol 8, 373 (2025).
27. Guan, X. et al. Predicting protein conformational motions using energetic frustration analysis and AlphaFold2. Proc Natl Acad Sci U S A 121, e2410662121 (2024).
28. Hariry, M. E., Cini, A., Mellone, G. & Balossino, A. Deep Reinforcement Learning Policies for Underactuated Satellite Attitude Control. https://doi.org/10.48550/ARXIV.2505.00165 (2025) doi:10.48550/ARXIV.2505.00165.
29. Xie, T., Song, Z. & Huang, J. Conditioned Protein Structure Prediction. PRX Life 2, 043001 (2024).
30. Chakravarty, D. et al. AlphaFold predictions of fold-switched conformations are driven by structure memorization. Nat Commun 15, 7296 (2024).
31. Chakravarty, D. & Porter, L. L. AlphaFold2 fails to predict protein fold switching. Protein Sci 31, e4353 (2022).
32. Schafer, J. W. & Porter, L. L. AlphaFold2's training set powers its predictions of some fold-switched conformations. Protein Sci 34, e70105 (2025).
33. Al-Masri, C. et al. Investigating the conformational landscape of AlphaFold2-predicted protein kinase structures. Bioinform Adv 3, vbad129 (2023).
34. Gizzio, J., Thakur, A., Haldane, A. & Levy, R. M. Evolutionary divergence in the conformational landscapes of tyrosine vs serine/threonine kinases. Elife 11, e83368 (2022).
35. Song, Z., Ding, Y. & Huang, J. Constant advance replicas method for locating minimum energy paths and transition states. J Comput Chem 44, 2042–2057 (2023).
36. Song, Z., Xu, Y., Zhang, H., Ding, Y. & Huang, J. Identifying the Most Probable Transition Path with Constant Advance Replicas. J Chem Theory Comput 20, 9857–9870 (2024).
37. Rosta, E., Nowotny, M., Yang, W. & Hummer, G. Catalytic mechanism of RNA backbone cleavage by ribonuclease H from quantum mechanics/molecular mechanics simulations. J Am Chem Soc 133, 8934–8941 (2011).
38. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. Sci Signal 12, eaau8645 (2019).
39. Hodgson, D. R. W. & Schröder, M. Chemical approaches towards unravelling kinase-mediated signalling pathways. Chem Soc Rev 40, 1211–1223 (2011).
40. Amano, M. et al. Kinase-interacting substrate screening is a novel method to identify kinase substrates. J Cell Biol 209, 895–912 (2015).
41. Obenauer, J. C., Cantley, L. C. & Yaffe, M. B. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31, 3635–3641 (2003).
42. Zhou, Z. et al. Using explainable machine learning to uncover the kinase-substrate interaction landscape. Bioinformatics 40, btae033 (2024).
43. Hornbeck, P. V. et al. PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res 40, D261-270 (2012).
44. Dinkel, H. et al. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res 39, D261-267 (2011).
45. Maier, B. D., Petursson, B., Lussana, A. & Petsalaki, E. Data-Driven Extraction of Human Kinase-Substrate Relationships From Omics Datasets. Mol Cell Proteomics 24, 100994 (2025).
46. Zhang, J., Cao, H., Gao, Z., Wang, X. & Gu, C. SAGEPhos: Sage Bio-Coupled and Augmented Fusion for Phosphorylation Site Detection. Preprint at https://doi.org/10.48550/arXiv.2502.07384 (2025).
47. Levinson, N. M. et al. A Src-like inactive conformation in the abl tyrosine kinase domain. PLoS Biol 4, e144 (2006).

672 48. Pettersen, E. F. et al. UCSF Chimera--a visualization system for exploratory research and
673 analysis. J Comput Chem 25, 1605–1612 (2004).
674 49. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. Proc Natl
675 Acad Sci U S A 89, 10915–10919 (1992).
676