Dark kinase annotation, mining, and visualization using the Protein Kinase Ontology

Saber Soleymani [1] , **Nathan Gravel** [2] , **Liang-Chin Huang** [2] , **Wayland Yeung** [2] , **Elika Bozorgi** [1] , **Nathaniel G Bendzunas** [3] , **Krzysztof J Kochut** [Corresp., 1] , **Natarajan Kannan** [Corresp. 2, 3]

[1] Department of Computer Science, University of Georgia, Athens, GA, United States

[2] Institute of Bioinformatics, University of Georgia, Athens, GA, United States

[3] Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, United States

Corresponding Authors: Krzysztof J Kochut, Natarajan Kannan
Email address: kkochut@uga.edu, nkannan@uga.edu

The Protein Kinase Ontology (ProKinO) is an integrated knowledge graph that conceptualizes the complex relationships among protein kinase sequence, structure, function, and disease in a human and machine-readable format. In this study, we have significantly expanded ProKinO by incorporating additional data on expression patterns and drug interactions. Furthermore, we have developed a completely new browser from the ground up to render the knowledge graph visible and interactive on the web. We have enriched ProKinO with new classes and relationships that capture information on kinase ligand binding sites, expression patterns, and functional features. These additions extend ProKinO's capabilities as a discovery tool, enabling it to uncover novel insights about understudied members of the protein kinase family. We next demonstrate the application of ProKinO. Specifically, through graph mining and aggregate SPARQL queries, we identify the p21-activated protein kinase 5 (PAK5) as one of the most frequently mutated dark kinases in human cancers with abnormal expression in multiple cancers, including a previously unappreciated role in acute myeloid leukemia. We have identified recurrent oncogenic mutations in the PAK5 activation loop predicted to alter substrate binding and phosphorylation. Additionally, we have identified common ligand/drug binding residues in PAK family kinases, underscoring ProKinO’s potential application in drug discovery. The updated ontology browser and the addition of a web component, ProtVista, which enables interactive mining of kinase sequence annotations in 3D structures and Alphafold models, provide a valuable resource for the signaling community. The updated ProKinO database is accessible at https://prokino.uga.edu.

Dark kinase annotation, mining, and visualization using the Protein Kinase Ontology

Saber Soleymani[1], Nathan Gravel[2], Liang-Chin Huang[2], Wayland Yeung[2], Elika Bozorgi[1], Nathaniel G. Bendzunas[3], Krzysztof J. Kochut[1] and Natarajan Kannan[2,3]

[1]Department of Computer Science, University of Georgia, Athens, GA 30602, USA
[2]Institute of Bioinformatics, University of Georgia, Athens, GA 30602, USA
[3]Department of Biochemistry & Molecular Biology, University of Georgia, Athens, GA 30602, USA

Corresponding Author:

Natarajan Kannan[2,3]
Department of Biochemistry and Molecular Biology, A318 Life Sciences, University of Georgia, Athens, GA, 30602, United States of America

Krzysztof J. Kochut[1]
School of Computing, 415 Boyd GSRC, University of Georgia, Athens, GA, 30602, United States of America

Email address:
nkannan@uga.edu
kkochut@uga.edu

Abstract

The Protein Kinase Ontology (ProKinO) is an integrated knowledge graph that conceptualizes the complex relationships among protein kinase sequence, structure, function, and disease in a human and machine-readable format. In this study, we have significantly expanded ProKinO by incorporating additional data on expression patterns and drug interactions. Furthermore, we have developed a completely new browser from the ground up to render the knowledge graph visible and interactive on the web.

We have enriched ProKinO with new classes and relationships that capture information on kinase ligand binding sites, expression patterns, and functional features. These additions extend ProKinO's capabilities as a discovery tool, enabling it to uncover novel insights about understudied members of the protein kinase family.

We next demonstrate the application of ProKinO. Specifically, through graph mining and aggregate SPARQL queries, we identify the p21-activated protein kinase 5 (PAK5) as one of the most frequently mutated dark kinases in human cancers with abnormal expression in multiple cancers, including a previously unappreciated role in acute myeloid leukemia. We have identified recurrent oncogenic mutations in the PAK5 activation loop predicted to alter substrate binding and phosphorylation. Additionally, we have identified common ligand/drug binding residues in PAK family kinases, underscoring ProKinO's potential application in drug discovery. The updated ontology browser and the addition of a web component, ProtVista, which enables interactive mining of kinase sequence annotations in 3D structures and Alphafold models, provide a valuable resource for the signaling community. The updated ProKinO database is accessible at https://prokino.uga.edu.

Introduction

The protein kinase gene family with nearly 535 human members (collectively called the human kinome) is a biomedically important gene family associated with many human diseases such as cancer, diabetes, Alzheimer's, Parkinson's, and inflammatory disorders. They make up one-third of all drug-related protein target discoveries in the pharmaceutical industry, with over 50 FDA-approved drugs developed since 2001 (Ferguson & Gray 2018; Zhang et al. 2009). However, despite decades of research on the protein kinase family, our current knowledge of the kinome is skewed towards a subset of well-studied kinases with nearly one-third of the kinome largely understudied. These understudied kinases, collectively referred to as the "dark" kinome by the Knowledge Management Center (KMC) (Nguyen et al. 2017) within the Illuminating the Druggable Genome (IDG) consortium, constitute both active kinases and inactive pseudokinases, which lack one or more of the active site residues, but perform important scaffolding and regulatory roles in signaling pathways (Byrne et al. 2017; Eyers et al. 2017; Eyers & Murphy 2013; Murphy et al. 2017) and are druggable (Foulkes et al. 2018). Incomplete knowledge of the structure, function, and regulation of these understudied kinases and pseudokinases presents a major bottleneck for drug discovery efforts. While multiple initiatives are beginning to generate essential tools and resources to characterize dark kinases, integrative mining of these datasets is necessary to develop new testable hypotheses on dark kinase functions. However, integrative mining of protein kinase data is a challenge because of the diverse and disparate nature of protein kinase data sources and formats. Information on the structural and functional aspects of dark kinases, for example, is scattered in the literature posing unique challenges for researchers interested in formulating routine queries such as “disease mutations mapping to conserved structural and functional regions of the kinome” or “post-translational modifications (PTMs) in the activation loop of dark kinases.” Formulating such aggregate queries requires researchers to go through the often time-consuming and error-prone process of collating information from various data sources through customized computer programs, which results in duplication of efforts across laboratories, and does not scale well with the growing complexity and diversity of protein kinase data. For these reasons, the IDG consortium has developed a unified resource, Pharos, for collating diverse forms of information on druggable proteins, including protein kinases (Nguyen et al. 2017; Sheils et al. 2020; Sheils et al. 2021). A focused Dark Kinase Knowledgebase has also been developed to make experimental data available on dark kinases to the broader research community (Berginski et al. 2021; Moret et al. 2021). While these unified resources provide a wide range of valuable information on druggable proteins, they offer limited data analytics capabilities in mining sequence and structural data. They do not conceptualize protein kinases' detailed structural and functional knowledge in a practical and understandable way for protein kinase researchers. Thus, to accelerate the biochemical characterization of understudied dark kinases, a semantically meaningful and mineable representation of the kinase knowledge base is needed (Fig. 1).

To semantically represent protein kinase data in ways protein kinase researchers use and understand, we previously reported the development of a focused protein kinase ontology, ProKinO (Gosal et al. 2011a; Gosal et al. 2011b; McSkimming et al. 2015), which integrates and conceptualizes diverse forms of protein kinase data in computer- and human-readable format (Fig. 2). The ontology is instantiated with curated data from internal and external sources and enables aggregate queries linking diverse forms of data in one place. ProKinO enables the generation of new knowledge regarding kinases and pathways altered in various cancer types, and new testable hypotheses regarding the structural and functional impact of disease mutations (Bailey et al. 2015; Cicenas & Cicenas 2016; Goldberg et al. 2013; Gosal et al. 2011a; Hu et al. 2015; Liu et al. 2016; McClendon et al. 2014; McSkimming et al. 2016; McSkimming et al. 2014; McSkimming et al. 2015; Meharena et al. 2013; Mohanty et al. 2016; Nguyen et al. 2015; Oruganty & Kannan 2013; Ruan & Kannan 2015; Simonetti et al. 2014; Taylor et al. 2015; U et al. 2014; Vazquez et al. 2016). For example, through iterative ProKinO queries and follow-up experimental studies, we identified oncogenic mutations associated with abnormal protein kinase activation and drug sensitivity (Lubner et al. 2017; McSkimming et al. 2016; McSkimming et al. 2015; Mohanty et al. 2016; Patani et al. 2016; Ruan & Kannan 2015; Ruan et al. 2017). We have also employed federated queries linking ProKinO with other widely used ontologies and resources such as the Protein Ontology (PRO), neXtProt, Reactome, and the Mouse Genome Informatics (MGI) to prioritize understudied dark kinases for functional studies and generate testable hypotheses regarding post-translational modification and cancer mutations (Huang et al. 2018).

While our preliminary studies have demonstrated the utility of ProKinO in hypothesis generation and knowledge discovery, to fully realize the impact of ProKinO in drug discovery and dark kinome mining, the ontology and the associated analytics tools need to be further developed to expand its scope and usability. For example, mutations at specific functional regions of the protein kinase domain, such as the gatekeeper and activation segments, are known to impact drug binding efficacies (Gajiwala et al. 2009; Yun et al. 2008). Likewise, kinase mRNA expression profiles strongly correlate with drug response (Benhar et al. 2002; Duncan et al. 2012; Kim et al. 2009; Niepel et al. 2017). Thus, integrative mining of disease mutations with drug sensitivity profiles and expression patterns can provide new hypotheses/data for the development and administration of combinatorial drugs where multiple mutated kinases in distinct pathways can be targeted for drug repurposing (Erika et al. 2016; Li & Jones 2012), as demonstrated by the repurposing of Gleevec for targeting c-kit kinase in Gastrointestinal tumors (Joensuu et al. 2001). Furthermore, the recent generation of structural models of various dark kinases using AlphaFold (Jumper et al. 2021) provides a new framework for generating new hypotheses by interactive mining and visualization of sequence annotations in the context of 3D models. However, the lack of interactive visualization tools to overlay sequence and functional annotations in 3D structural models presents a bottleneck in the effective use of AlphaFold models for function prediction. To address this and other challenges described above related to dark kinase mining and annotation, we have expanded ProKinO by including kinase expression data, as well as a variety of data related to ligand-motif interaction, and ligand response prediction (Huang et al. 2021). We have also significantly revamped the ProKinO browser through incorporation of new visualization tools for the interactive mining of sequence annotations in the context of experimentally determined 3D structures and AlphaFold models. We demonstrate the application of these new tools in dark kinase annotation and mining using the understudied p21-activated protein kinase 5, as an example. The updated ontology and browser provide a valuable resource for mining, visualizing, and annotating the dark kinome and pseudokinome.

Materials & Methods

Data Sources.

The ProKinO ontology includes data obtained from curated internal sources as well as external sources. Information from internal sources include annotations of kinase sequence and structural motifs retrieved from curated multiple sequence alignments. External sources are used for information related to kinase sequence and classification (KinBase & UniProt) (Bairoch et al. 2005; Manning et al. 2002) cancer mutations (COSMIC) (Tate et al. 2018), pathways (Reactome) (Croft et al. 2011) and three dimernsioanl structure (PDB) (Berman et al. 2000). The ontology is populated and updated on a regular basis using protocols described in previous studies (Gosal et al. 2011b; McSkimming et al. 2016). Here, we describe further enhancements and additions to ProKinO through integration of data on kinase expression patterns and drug interactions, as described below. In a seperate significant project, we have identified and classified nearly 30,000 pseudokinases spanning over 1,300 organisms (Kwon et al. 2019). The schematic representation of the classification of kinases into groups, families, and subfamilies was already in place (Hanks & Hunter 1995; Manning et al. 2002). Consequently, the addition of the pseudokinases and their classification was relatively simple. However, it significantly enhanced ProKinO as a comprehensive knowledge graph representing kinase-related data. The definition and nomenclature of several kinome-wide conserved motifs were standardized based on several previously published studies which describe the kinase structural features such as subdomains (Hanks & Hunter 1995), regulatory spine/shell (Meharena et al. 2013), and catalytic spine (Hu et al. 2015). A subset of redundant or family-specific motifs were removed in the updated ontology and motif information on some of the atypical kinases such as ALPHAK2 is not included as they cannot be reliably aligned with canonical protein kinases. Portions of the text were previously published as a part of a preprint (Soleymani et al. 2022) (https://www.biorxiv.org/content/biorxiv/early/2022/03/01/2022.02.25.482021.full.pdf).

Ligand interactions.

Information on kinase-ligand interactions were retrieved from the Kinase-Ligand Interaction Fingerprints and Structures (KLIFS) database (Kanev et al. 2021). The KLIFS database stores detailed drug-protein kinase interaction information derived from diverse (>2900) structures of catalytic domains of human and mouse protein kinases deposited in the Protein Data Bank. In addition, KLIFS provides an Application Programming Interface (API) for programmatic access to data related to chemicals and structural chemogenomics (Kanev et al. 2021). However, it lacks information regarding kinase pathways or diseases which prevents the user from investigating the effect of drug-mutant protein binding on downstream pathways or diseases. KLIFS annotations, which report PDB residue positions, were converted to UniProt residue numbering using PDBrenum (Faezov & Dunbrack 2021), then converted to prototypic Protein Kinase A (PKA) numbering using Multiply Aligned Profiles for Global Alignment of Protein Sequences (MAPGAPS) (Neuwald 2009). Entries that could not be mapped or did not map to the kinase domain were filtered out.

Ligand responses.

We have also incorporated information on kinase drug sensitivity profile in the updated ProKinO. In particular, we retrieved drug dose response data for kinase-relevant ligands/drugs from the Genomics of Drug Sensitivity in Cancer (GDSC) (Yang et al. 2013). Kinase-relevant ligands are defined based on our previous study (Huang et al. 2020), which collected 143 small-molecule protein kinase inhibitors from GDSC based on four drug-target databases: DrugBank (Wishart et al. 2018), Therapeutic Target Database (Li et al. 2018), Pharos (Nguyen et al. 2017), and The Library of Integrated Network-Based Cellular Signatures (LINCS) Data Portal (Koleti et al. 2018). GDSC provides the half-maximal inhibitory concentration values (IC50) of these 143 ligands in 988 cancer cell lines.

Ligand activities.

Ligand activities were retrieved from Pharos, a flagship resource (Nguyen et al. 2017) of the National Institutes of Health (NIH) Illuminating the Druggable Genome (IDG) program that includes data on small molecules, including approved drug data and bioassay data. Based on the protein classification (Lin et al. 2017), the drug targets in Pharos include kinases, ion channels, G-protein coupled receptors (GPCRs), and others. In this phase of the project, we decided to include the data relevant to ligand binding in kinases. Pharos integrates drug-target relationships from several resources, such as ChEMBL (Bühlmann & Reymond 2020) and DrugCentral (Avram et al. 2021).

Expression data.

An important part of our recent additions was kinase expression data. Genomic expression data (protein, RNA), as well as transcription factors and epigenomic associations, are among many facets of the data included in Pharos. Furthermore, the GDSC repository contains gene expression data (Affymetrix Human Genome U219 Array), as well. Additionally, COSMIC's Cell Lines Project includes a significant amount of gene expression data, including kinase expression.

Dark kinases.

Dark kinases were labeled based on the information from Dark Kinase Knowledgebase (Berginski et al. 2021).

Protein kinase knowledge graph: schema and data organization.

The ProKinO ontology consists of classes, sub-classes, class types, relationships, relationship types, and constraints of protein kinase and related data (Fig. 2). The hierarchy connects all classes to the root, which is ProKinOEntity. Moreover, the schema defines types and constraints for the relationships. With such explicit and constrained schema, composing queries is more intuitive than conventional relational databases. In particular, to enable integrative mining of dark kinase expression data in the context of kinase sequence and structural features, we have introduced three new classes in ProKinO, the Ligand class (including its name, source, and chemical structure) and the following three related classes: (1) LigandInteraction, placed between the Ligand and (already existing) Motif classes to capture kinase-ligand binding and selectivity at the motif and residue level, (2) LigandActivity, placed between the Ligand and (already existing) Protein classes to represent kinases targeted by ligands (and drugs), and (3) LigandResponse, located between the Ligand and (already existing) Sample classes and representing ligand (and drug) sensitivity in kinases. To capture kinase expression, we added the GeneExpression relationship linking the Protein and Sample classes. The outline of the recently added classes and their relationships in ProKinO is illustrated as a UML class diagram, shown in Figure 2.

ProKinO population.

The ProKinO knowledge graph is automatically populated from several external and local data sources at regular intervals, as originally described (Gosal et al. 2011b), ProKinO schema and the associated knowledge graph population software are routinely updated to incorporate additional sources of data such as pseudokinase and "dark" kinase classification and incorporating information on ligand interactions, ligand responses, ligand activities, kinase expression and associated object and datatype properties. We have been using the Protégé ontology editor for the schema creation and its subsequent modifications. The organization of the schema after these modifications is available at https://prokino.uga.edu/about.

The population software has been coded in Java and uses the Jena Framework. The population process is performed in several steps to add instances, their properties, and a combination of reading the prepared data from CSV, RDF, XML, and other file formats and accessing many remote data sources using their provided API (for example, Reactome's REST API). Entity interconnections across data retrieved from different data sources are accomplished using UniProt identifiers, kinase names, and other accession identifiers. We modified the population software to create instances and properties for the newly added classes and relationships.

More specifically, using the KLIFS API, we retrieved the relevant kinases, ligands, and residue-level interaction data. The data was retrieved and then processed by custom Perl scripts. ProKinO ontology schema was modified, and ligands were included as new data, while interaction data (motifs) were either reconciled with the motifs already present in ProKinO or added as new, if not already there.

Similarly, the ligand response data was retrieved from GDSC and then processed by custom Perl scripts to create suitable CSV files. Additional ligands were included as new data, while the response data and the relevant samples were either reconciled with the samples already present in ProKinO or added as new, if not already there.

In order to populate the data on ligand activities, we retrieved from Pharos kinase-relevant ligands, as well as their binding data on targeted kinases, for example, IC50 values. This data was retrieved and then processed by custom Perl scripts to produce the necessary CSV files. Additional ligands, not included in the KLIFS dataset, were included as new data. All kinases targeted by ligands were already present in ProKinO, so they were reused in this step.

Data on kinase expression was first retrieved from Pharos, COSMIC, and GDSC. As before, the relevant kinases were already present in the ProKinO knowledge graph. The expression data was stored as individuals in the Expression class. Some of the relevant data about samples were already present in ProKinO, as we already had a significant amount of sample data from COSMIC. Additional samples were included as new data.

We reviewed and updated all the motifs already present in ProKinO. Furthermore, we updated the motif naming in cases where there were differences with the standard motif names.

Finally, we assembled an up-to-date list of dark kinases (Berginski et al. 2021) and added a Boolean datatype property, isDarkKinase, to identify them among all other kinases in the ProKinO knowledge graph.

Results / Discussion

The expanded ontology and its knowledge graph provide a wealth of data unifying the information available on both well-studied (light) kinases and understudied (dark) kinases that serve as a unified resource for mining the kinome. The current version of ProKinO (version 65), includes 842 classes, 31 objects and 67 data properties, and over seven million individuals (knowledge graph nodes). ProKinO contains information on 153 dark kinases. 137 dark kinases have information on structural motifs, 148 have disease mutations mapped to the kinase domain, 45 dark kinases have pathway information, and 26 are associated with specific reactions, as defined in Reactome.

Users can navigate the ontology using the ontology browser by searching for a specific kinase of interest or by performing aggregate SPARQL queries linking multiple forms of data. Currently 35 pre-written queries linking different data types can be executed using the ProKinO browser (http://prokino.uga.edu/queries). A user can also download the ontology or browse data based on organisms, functional domains, diseases, or kinase domain evolutionary hierarchy. Below, we focus on the application of complex SPARQL queries and the ProtVista visualization tools for the illumination of understudied dark kinases.

Mutation and expression of understudied PAK5 in human cancers.

One possible way to prioritize dark kinases for functional studies is to ask the question, “which dark kinases are most mutated in human diseases, such as cancers?”. Typically answering this question would require collating and post-processing data from multiple resources such as COSMIC, Pharos, and the Dark Kinase Knowledgebase. However, with the updated Protein Kinase Ontology, these questions can be quickly answered using SPARQL. Having the “*isDarkKinase”* property within the Protein class and the RDF triples connecting the “*Mutation”*, “*Sample*” and “*Sequence”* classes, one can formulate aggregate queries requesting all dark kinases mutated in cancer samples. To avoid biases introduced by the length of protein/gene sequences (longer proteins tend to have more mutations), the query can be modified to normalize mutation counts by sequence length. Executing this modified query (Query 27, available at http://prokino.uga.edu/queries) displays the rank-ordered list of dark kinases based on mutational density. The top ten dark kinases with the highest mutational density are shown in Figure 3A. Notably, the p21 activated kinase 5 (PAK5) is at the top of the list with a mutational density of 1.917, followed by CRK7 (1.054), PKACG (1.011), PSKH2 (1.01) TSSK1 (1.008), CK1A2 (0.991), ERK4 (0.966), DCLK3 (0.912), PKCT (0.894) and PAK3 (0.859). Having identified PAK5 as the most frequently mutated dark kinase in cancers, one can further query the ontology to explore the role of this kinase in various cancers. With the addition of the new "*GeneExpression*" class in ProKinO and the RDF triples connecting gene expression to the "*Sample"* and "P*rotein*" classes (*GeneExpression:InSample: Sample; GeneExpression:hasProtein: Protein*), one can formulate queries for PAK5 expression in different samples (Fig. 3B). Rank ordering the samples based on PAK5 expression (Query 33) reveals cancer types such as adenocarcinoma (Zscore: 4701.5) and hepatocellular carcinoma (Zscore: 2038.3) that have previously been associated with abnormal PAK5 expression (Fang et al. 2014; Han et al. 2018; Huo et al. 2019; Zhang et al. 2017). However, the role of PAK5 in other cancer types such as acute myeloid leukemia (Zscore: 136.5) is relatively underappreciated (Quan et al. 2020). The identification of new cancer sub-types with dark kinase expression and regulation further exemplifies the use of ProKinO in knowledge discovery.

Mutational hotspots in the activation loop of PAK5.

Because ProKinO encodes a wealth of information on the structural and regulatory properties of multiple kinases, it can be used to generate mechanistic predictions on cancer mutation impact. We demonstrate this for the PAK kinases by asking the question "*where are PAK5 mutations located in the protein kinase domain?*" Using the RDF triples connecting the "*Mutation"*, "*Motif"* and "*Sequence*" classes ("*Mutation: LocatedIn: Motif"; "Mutation:InSequence: Sequence"*), one can formulate a query (Query 28) listing mutations in different structural regions/motifs of the PAK5 kinase domain. Examination of the query results reveals that the C-terminal substrate binding lobe (C-lobe) is more frequently mutated (320 mutations) relative to the N-terminal ATP binding lobe (N-lobe: 173 mutations) (Fig. 4A). Within the C-lobe, nearly 78 mutations map to the activation loop, which is known to play a critical role in substrate recognition and activation in a diverse array of kinases (Huse & Kuriyan 2002; Kornev & Taylor 2015; Oruganty & Kannan 2012). Despite the prevalence of activation loop mutations in PAK5, there is currently no information on how these mutations impact PAK5 kinase structure and function. Nonetheless, based on the evolutionary relationships captured in ProKinO (based on the alignment of human kinases to the prototypic protein kinase A), one can formulate queries mapping mutations to specific aligned positions in the shared protein kinase domain. A query listing (wild type) WT type and mutant type residues in the activation loop of PAK5 and the equivalent aligned residue positions in PKA (Query 29) provides additional context for these mutations. For example, two distinct mutations map to residue P602$^{\text{PAK5}}$ in the activation loop of PAK5 that structurally corresponds to a phosphorylatable residue, T197$^{\text{PKA}}$, in PKA (Yonemoto et al. 1993). Having this context provides a testable hypothesis that S602 mutations in PAK5 impact kinase phosphorylation and regulation. Likewise, WT residue P607$^{\text{PAK5}}$ is mutated in four distinct cancer samples and this position is equivalent to PKA residue P202$^{\text{PKA}}$, which configures the activation loop for substrate recognition (Knighton et al. 1991). Thus, mutation of this critical residue is expected to impact substrate binding and activation loop phosphorylation in PAK5. Additional insights into these mutations can also be obtained by visualizing these residues in the context of the PAK5 AlphaFold models using the ProtVista viewer described below.

Insights into PAK5 ligand binding sites.

With the conceptualization of new information related to kinase ligands, their mode of action and interaction with specific motifs in the kinase domain, new aggregate queries linking mutated kinases to drug sensitivity profiles, mode of action, and ligand binding sites can be performed using the updated ProKinO. For example, queries such as "*list proteins and drugs or ligands interacting with the protein's gatekeeper residue (GK.45)*" (Query 31) and "*list ligands targeting the Epidermal Growth Factor Receptor (EGFR) kinase and their mode of action*" (Query 34) can be rapidly performed using the updated ProKinO ontology. We demonstrate the application of these new additions in the context of PAK5 by asking the question "*what are the drugs targeting PAK family (PAK1-6) kinases*?" Query 30 answers this question using the RDF triples connecting the "*Ligand*", "*Motif*" and "*Protein*" classes (list triples) (Fig. 5). Examination of the query results indicates multiple drugs targeting PAK family kinases, including STAUROSPORINE and N2-[(1R-2S)-2-AMINOCYCLOHEXYL] that bind to structurally equivalent residues/motifs in the ligand binding pocket of PAK4 and PAK5, respectively. The ligand binding sites, and associated interactions can also be visualized using the ProtVista viewer described below. Additional queries linking dark kinases to drug sensitivities, structural motifs, and pathways are listed on the ProKinO website at https://prokino.uga.edu/queries.

Visualization tools for dark kinase annotation and mining.

To provide structural context for cancer mutations and to enable interactive mining of dark kinase sequence annotations in the context of 3D structures and predicted models from AlphaFold (Jumper et al. 2021; Tunyasuvunakool et al. 2021), we developed and incorporated a modified version of the ProtVista viewer in ProKinO. The viewer can be deployed for any protein kinase of interest by navigating to the Structure tab in the protein summary page and selecting either a PDB structure or AlphaFold model of interest. A snapshot of the ProtVista viewer displaying the AlphaFold model of PAK5 kinase is shown in Figure 6. The ProtVista viewer uses an enhanced version of the Mol* viewer and the PDB web component (Watkins et al. 2017) to provide two-way interactive navigation between the 3D structure (Fig. 6A, top panel) and annotation viewer (Fig. 6A, bottom panel).

The annotation viewer consists of multiple tracks populated dynamically based on data from ProKinO and external sources such as UniProt. In addition, prediction confidence scores for AlphaFold models are displayed in the annotation viewer along with additional annotations such as conserved sequence motifs, subdomains, and structural motifs involved in kinase regulation. The annotation viewer also shows other annotations from external sources such as ligand binding sites and predicted functional sites. Users can hover over the residues on the 3D structure viewer to view the equivalent information on the annotation viewer and vice versa. For example, selecting the "*activation loop*" in the annotation viewer highlights the corresponding structural region in the AlphaFold model of PAK5 (Fig. 6A). Likewise, the selection of residues in the activation loop (S602 and P607) in the structure viewer highlights the annotations associated with these and interacting residues in the sequence viewer. Such interactive mining is expected to accelerate the functional characterization of dark kinases and provide new insights into disease mutations. For example, visualizing the interactions associated with S602 in the activation loop of PAK5 (Fig. 6B) indicates a hydrogen bonding interaction with R567, which is part of the conserved HRD motif (sequence annotation). Because the HRD-Arg is known to play a role in kinase regulation by stabilizing activation loop conformation (Huse & Kuriyan 2002), it provides additional context for predicting the impact of S602 altering mutations. Likewise, examining the structural and sequence context of P607 interacting residues provides new insights into how the alteration of this residue might impact substrate binding and kinase regulation. Together, these examples highlight the value added by the ProtVista viewer in the visualization and annotation of mutations in dark kinases.

Conclusions

This work presents an updated version of the Protein Kinase Otology (ProKinO) for mining and annotating dark kinases. ProKinO was developed following FAIR (Findable, Accessible, Interoperable, and Reusable) principles (Wilkinson et al. 2016) and serves as an integrated knowledge graph for relating and conceptualizing diverse forms of disparate data related to protein kinase sequence, structure, function, regulation, and disease (cancer). We present a new ontology browser for navigating these data and demonstrate the application of aggregate SPARQL queries in uncovering new testable hypotheses regarding understudied kinase members. We also provide several pre-written SPARQL queries that can rapidly retrieve information related to protein kinase mutations, pathways, expression, and ligand binding sites. However, writing new queries requires prior knowledge of the ontology schema and the SPARQL query language, which most bench biologists may not have. To alleviate this challenge, we are currently building a graphical SPARQL query interface, which will intuitively enable query formulation through the navigation of the knowledge graph schema. We are also exploring the application of ProKinO for machine learning-based knowledge discovery and hypotheses generation.

Acknowledgements

We acknowledge members of the Kannan lab for their valuable comments and suggestions. We additionally acknowledge the various contributions to the databases we utilized made through the efforts of the IDG consortium and numerous investigator-initiated efforts.

Funding.

This work was supported by the National Institutes of Health (U01CA239106 and R35 GM139656) to NK. The funding bodies did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Competing Interests.

The authors declare that they have no competing interests.

Author Contributions.

NK conceived the project. SS, EB, and KK updated the browser and ontology. SS, KK, NG, LH, and WY patriated in data curation for the updated the ontology. NG, SS, KK, and NK designed the experiments, analyzed the data, interpreted the results, and visualized the data. SS, NG, KK, and NK wrote the manuscript. SS, NG, NB, KK, and NK revised the manuscript. All authors read and approved the final manuscript.

Data Availability.

The protein kinase ontology (ProKinO)'s latest OWL file and previous versions are publicly available at https://prokino.uga.edu/downloads.html. Future versions of the ontology also will be placed at the same address. Also, the ontology browser is accessible at https://prokino.uga.edu/browser. Users can save the results of queries in diagrams or other formats such as CSV.

References

Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, Curpan R, Halip L, Bora A, Yang JJ, Knockel J, Sirimulla S, Ursu O, and Oprea TI. 2021. DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res* 49:D1160-d1169. 10.1093/nar/gkaa997

Bailey FP, Byrne DP, McSkimming D, Kannan N, and Eyers PA. 2015. Going for broke: targeting the human cancer pseudokinome. *Biochem J* 465:195-211. 10.1042/BJ20141060

Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, and Magrane M. 2005. The universal protein resource (UniProt). *Nucleic Acids Research* 33:D154-D159.

Benhar M, Engelberg D, and Levitzki A. 2002. ROS, stress-activated kinases and stress signaling in cancer. *EMBO reports* 3:420-425.

Berginski ME, Moret N, Liu C, Goldfarb D, Sorger PK, and Gomez SM. 2021. The Dark Kinase Knowledgebase: an online compendium of knowledge and experimental results of understudied kinases. *Nucleic Acids Res* 49:D529-d535. 10.1093/nar/gkaa853

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE. 2000. The protein data bank. *Nucleic Acids Research* 28:235-242.

Bühlmann S, and Reymond JL. 2020. ChEMBL-Likeness Score and Database GDBChEMBL. *Front Chem* 8:46. 10.3389/fchem.2020.00046

Byrne DP, Foulkes DM, and Eyers PA. 2017. Pseudokinases: update on their functions and evaluation as new drug targets. *Future Med Chem* 9:245-265. 10.4155/fmc-2016-0207

Cicenas J, and Cicenas E. 2016. Multi-kinase inhibitors, AURKs and cancer. *Med Oncol* 33:43. 10.1007/s12032-016-0758-4

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, and Jassal B. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* 39:D691-D697.

Duncan JS, Whittle MC, Nakamura K, Abell AN, Midland AA, Zawistowski JS, Johnson NL, Granger DA, Jordan NV, Darr DB, Usary J, Kuan PF, Smalley DM, Major B, He X, Hoadley KA, Zhou B, Sharpless NE, Perou CM, Kim WY, Gomez SM, Chen X, Jin J, Frye SV, Earp HS, Graves LM, and Johnson GL. 2012. Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. *Cell* 149:307-321. 10.1016/j.cell.2012.02.053

Erika G, Federica Z, Martina S, Anselmo P, Luigi R, Marina M, Davide C, Eleonora Z, Monica V, and Silverio T. 2016. Old Tyrosine Kinase Inhibitors and Newcomers in Gastrointestinal Cancer Treatment. *Curr Cancer Drug Targets* 16:175-185.

Eyers PA, Keeshan K, and Kannan N. 2017. Tribbles in the 21st Century: The Evolving Roles of Tribbles Pseudokinases in Biology and Disease. *Trends Cell Biol* 27:284-298. 10.1016/j.tcb.2016.11.002

Eyers PA, and Murphy JM. 2013. Dawn of the dead: protein pseudokinases signal new adventures in cell biology. *Biochem Soc Trans* 41:969-974. 10.1042/BST20130115

Faezov B, and Dunbrack RL, Jr. 2021. PDBrenum: A webserver and program providing Protein Data Bank files renumbered according to their UniProt sequences. *PLoS One* 16:e0253411. 10.1371/journal.pone.0253411

Fang ZP, Jiang BG, Gu XF, Zhao B, Ge RL, and Zhang FB. 2014. P21-activated kinase 5 plays essential roles in the proliferation and tumorigenicity of human hepatocellular carcinoma. *Acta Pharmacol Sin* 35:82-88. 10.1038/aps.2013.31

Ferguson FM, and Gray NS. 2018. Kinase inhibitors: the road ahead. *Nat Rev Drug Discov* 17:353-377. 10.1038/nrd.2018.21

Foulkes DM, Byrne DP, Yeung W, Shrestha S, Bailey FP, Ferries S, Eyers CE, Keeshan K, Wells C, and Drewry DH. 2018. Covalent inhibitors of EGFR family protein kinases induce degradation of human Tribbles 2 (TRIB2) pseudokinase in cancer cells. *Science signaling* 11:eaat7951.

Gajiwala KS, Wu JC, Christensen J, Deshmukh GD, Diehl W, DiNitto JP, English JM, Greig MJ, He Y-A, and Jacques SL. 2009. KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proceedings of the National Academy of Sciences* 106:1542-1547.

Goldberg JM, Griggs AD, Smith JL, Haas BJ, Wortman JR, and Zeng Q. 2013. Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. *Bioinformatics* 29:2387-2394. 10.1093/bioinformatics/btt419

Gosal G, Kannan N, and Kochut K. 2011a. ProKinO: A framework for protein kinase ontology. *Proceedings of the IEEE International Conference on Bioinformatics & Biomedicine, Atlanta, Georgia*:550-555.

Gosal G, Kochut KJ, and Kannan N. 2011b. ProKinO: an ontology for integrative analysis of protein kinases in cancer. *PLoS One* 6:e28782. 10.1371/journal.pone.0028782
PONE-D-11-14617 [pii]

Han K, Zhou Y, Tseng KF, Hu H, Li K, Wang Y, Gan Z, Lin S, Sun Y, and Min D. 2018. PAK5 overexpression is associated with lung metastasis in osteosarcoma. *Oncol Lett* 15:2202-2210. 10.3892/ol.2017.7545

Hanks SK, and Hunter T. 1995. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *Faseb j* 9:576-596.

Hu J, Ahuja LG, Meharena HS, Kannan N, Kornev AP, Taylor SS, and Shaw AS. 2015. Kinase regulation by hydrophobic spine assembly in cancer. *Mol Cell Biol* 35:264-276. 10.1128/MCB.00943-14

Huang LC, Ross KE, Baffi TR, Drabkin H, Kochut KJ, Ruan Z, D'Eustachio P, McSkimming D, Arighi C, Chen C, Natale DA, Smith C, Gaudet P, Newton AC, Wu C, and Kannan N. 2018. Integrative annotation and knowledge discovery of kinase post-translational modifications and cancer-associated mutations through federated protein ontologies and resources. *Sci Rep* 8:6518. 10.1038/s41598-018-24457-1

Huang LC, Taujale R, Gravel N, Venkat A, Yeung W, Byrne DP, Eyers PA, and Kannan N. 2021. KinOrtho: a method for mapping human kinase orthologs across the tree of life and illuminating understudied kinases. *BMC Bioinformatics* 22:446. 10.1186/s12859-021-04358-3

Huang LC, Yeung W, Wang Y, Cheng H, Venkat A, Li S, Ma P, Rasheed K, and Kannan N. 2020. Quantitative Structure-Mutation-Activity Relationship Tests (QSMART) model for protein kinase inhibitor response prediction. *BMC Bioinformatics* 21:520. 10.1186/s12859-020-03842-6

Huo FC, Pan YJ, Li TT, Mou J, and Pei DS. 2019. PAK5 promotes the migration and invasion of cervical cancer cells by phosphorylating SATB1. *Cell Death Differ* 26:994-1006. 10.1038/s41418-018-0178-4

Huse M, and Kuriyan J. 2002. The conformational plasticity of protein kinases. *Cell* 109:275-282. 10.1016/s0092-8674(02)00741-9

Joensuu H, Roberts PJ, Sarlomo-Rikala M, Andersson LC, Tervahartiala P, Tuveson D, Silberman S, Capdeville R, Dimitrijevic S, Druker B, and Demetri GD. 2001. Effect of the tyrosine kinase inhibitor STI571 in a patient with a metastatic gastrointestinal stromal tumor. *N Engl J Med* 344:1052-1056. 10.1056/NEJM200104053441404

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 SAA, Ballard AJ, 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 AW, Kavukcuoglu K, Kohli P, and Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583-589. 10.1038/s41586-021-03819-2

Kanev GK, de Graaf C, Westerman BA, de Esch IJP, and Kooistra AJ. 2021. KLIFS: an overhaul after the first 5 years of supporting kinase research. *Nucleic Acids Res* 49:D562-d569. 10.1093/nar/gkaa895

Kim LC, Song L, and Haura EB. 2009. Src kinases as therapeutic targets for cancer. *Nature Reviews Clinical Oncology* 6:587-595.

Knighton DR, Zheng JH, Ten Eyck LF, Ashford VA, Xuong NH, Taylor SS, and Sowadski JM. 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253:407-414. 10.1126/science.1862342

Koleti A, Terryn R, Stathias V, Chung C, Cooper DJ, Turner JP, Vidovic D, Forlin M, Kelley TT, D'Urso A, Allen BK, Torre D, Jagodnik KM, Wang L, Jenkins SL, Mader C, Niu W, Fazel M, Mahi N, Pilarczyk M, Clark N, Shamsaei B, Meller J, Vasiliauskas J, Reichard J, Medvedovic M, Ma'ayan A, Pillai A, and Schürer SC. 2018. Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. *Nucleic Acids Res* 46:D558-d566. 10.1093/nar/gkx1063

Kornev AP, and Taylor SS. 2015. Dynamics-Driven Allostery in Protein Kinases. *Trends Biochem Sci* 40:628-647. 10.1016/j.tibs.2015.09.002

Kwon A, Scott S, Taujale R, Yeung W, Kochut KJ, Eyers PA, and Kannan N. 2019. Tracing the origin and evolution of pseudokinases across the tree of life. *Sci Signal* 12. 10.1126/scisignal.aav3810

Li YH, Yu CY, Li XX, Zhang P, Tang J, Yang Q, Fu T, Zhang X, Cui X, Tu G, Zhang Y, Li S, Yang F, Sun Q, Qin C, Zeng X, Chen Z, Chen YZ, and Zhu F. 2018. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 46:D1121-d1127. 10.1093/nar/gkx1076

Li YY, and Jones SJ. 2012. Drug repositioning for personalized medicine. *Genome Med* 4:27. 10.1186/gm326

Lin Y, Mehta S, Küçük-McGinty H, Turner JP, Vidovic D, Forlin M, Koleti A, Nguyen DT, Jensen LJ, Guha R, Mathias SL, Ursu O, Stathias V, Duan J, Nabizadeh N, Chung C, Mader C, Visser U, Yang JJ, Bologa CG, Oprea TI, and Schürer SC. 2017. Drug target ontology to classify and integrate drug discovery data. *J Biomed Semantics* 8:50. 10.1186/s13326-017-0161-x

Liu D, Liang XC, and Zhang H. 2016. Culturing Schwann Cells from Neonatal Rats by Improved Enzyme Digestion Combined with Explants-culture Method. *Zhongguo Yi Xue Ke Xue Yuan Xue Bao* 38:388-392. 10.3881/j.issn.1000-503X.2016.04.004

Lubner JM, Dodge-Kafka KL, Carlson CR, Church GM, Chou MF, and Schwartz D. 2017. Cushing's syndrome mutant PKA(L)(205R) exhibits altered substrate specificity. *FEBS Lett* 591:459-467. 10.1002/1873-3468.12562

Manning G, Whyte DB, Martinez R, Hunter T, and Sudarsanam S. 2002. The protein kinase complement of the human genome. *Science* 298:1912-1934.

McClendon CL, Kornev AP, Gilson MK, and Taylor SS. 2014. Dynamic architecture of a protein kinase. *Proc Natl Acad Sci U S A* 111:E4623-4631. 10.1073/pnas.1418402111

McSkimming DI, Dastgheib S, Baffi TR, Byrne DP, Ferries S, Scott ST, Newton AC, Eyers CE, Kochut KJ, Eyers PA, and Kannan N. 2016. KinView: a visual comparative sequence analysis tool for integrated kinome research. *Mol Biosyst*. 10.1039/c6mb00466k

McSkimming DI, Dastgheib S, Talevich E, Narayanan A, Katiyar S, Taylor SS, Kochut K, and Kannan N. 2014. ProKinO: A Unified Resource for Mining the Cancer Kinome. *Hum Mutat*. 10.1002/humu.22726

McSkimming DI, Dastgheib S, Talevich E, Narayanan A, Katiyar S, Taylor SS, Kochut K, and Kannan N. 2015. ProKinO: a unified resource for mining the cancer kinome. *Hum Mutat* 36:175-186. 10.1002/humu.22726

Meharena HS, Chang P, Keshwani MM, Oruganty K, Nene AK, Kannan N, Taylor SS, and Kornev AP. 2013. Deciphering the structural basis of eukaryotic protein kinase regulation. *PLoS Biol* 11:e1001680. 10.1371/journal.pbio.1001680 PBIOLOGY-D-13-02013 [pii]

Mohanty S, Oruganty K, Kwon A, Byrne DP, Ferries S, Ruan Z, Hanold LE, Katiyar S, Kennedy EJ, Eyers PA, and Kannan N. 2016. Hydrophobic Core Variations Provide a Structural Framework for Tyrosine Kinase Evolution and Functional Specialization. *PLoS Genet* 12:e1005885. 10.1371/journal.pgen.1005885

Moret N, Liu C, Gyori BM, Bachman JA, Steppi A, Hug C, Taujale R, Huang L-C, Berginski ME, and Gomez SM. 2021. A resource for exploring the understudied human kinome for research and therapeutic opportunities. *BioRxiv*:2020.2004. 2002.022277.

Murphy JM, Mace PD, and Eyers PA. 2017. Live and let die: insights into pseudoenzyme mechanisms from structure. *Curr Opin Struct Biol* 47:95-104. 10.1016/j.sbi.2017.07.004

Neuwald AF. 2009. Rapid detection, classification and accurate alignment of up to a million or more related protein sequences. *Bioinformatics* 25:1869-1875. btp342 [pii] 10.1093/bioinformatics/btp342

Nguyen DT, Mathias S, Bologa C, Brunak S, Fernandez N, Gaulton A, Hersey A, Holmes J, Jensen LJ, Karlsson A, Liu G, Ma'ayan A, Mandava G, Mani S, Mehta S, Overington J, Patel J, Rouillard AD, Schurer S, Sheils T, Simeonov A, Sklar LA, Southall N, Ursu O, Vidovic D, Waller A, Yang J, Jadhav A, Oprea TI, and Guha R. 2017. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res* 45:D995-D1002. 10.1093/nar/gkw1072

Nguyen T, Ruan Z, Oruganty K, and Kannan N. 2015. Co-conserved MAPK features couple D-domain docking groove to distal allosteric sites via the C-terminal flanking tail. *PLoS One* 10:e0119636. 10.1371/journal.pone.0119636

Niepel M, Hafner M, Duan Q, Wang Z, Paull EO, Chung M, Lu X, Stuart JM, Golub TR, Subramanian A, Ma'ayan A, and Sorger PK. 2017. Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling. *Nat Commun* 8:1186. 10.1038/s41467-017-01383-w

Oruganty K, and Kannan N. 2012. Design principles underpinning the regulatory diversity of protein kinases. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367:2529-2539.

Oruganty K, and Kannan N. 2013. Evolutionary variation and adaptation in a conserved protein kinase allosteric network: Implications for inhibitor design. *Biochim Biophys Acta*. S1570-9639(13)00111-8 [pii]
10.1016/j.bbapap.2013.02.040

Patani H, Bunney TD, Thiyagarajan N, Norman RA, Ogg D, Breed J, Ashford P, Potterton A, Edwards M, Williams SV, Thomson GS, Pang CS, Knowles MA, Breeze AL, Orengo C, Phillips C, and Katan M. 2016. Landscape of activating cancer mutations in FGFR kinases and their differential responses to inhibitors in clinical use. *Oncotarget* 7:24252-24268. 10.18632/oncotarget.8132

Quan L, Cheng Z, Dai Y, Jiao Y, Shi J, and Fu L. 2020. Prognostic significance of PAK family kinases in acute myeloid leukemia. *Cancer Gene Ther* 27:30-37. 10.1038/s41417-019-0090-1

Ruan Z, and Kannan N. 2015. Mechanistic Insights into R776H Mediated Activation of Epidermal Growth Factor Receptor Kinase. *Biochemistry* 54:4216-4225. 10.1021/acs.biochem.5b00444

Ruan Z, Katiyar S, and Kannan N. 2017. Computational and Experimental Characterization of Patient Derived Mutations Reveal an Unusual Mode of Regulatory Spine Assembly and Drug Sensitivity in EGFR Kinase. *Biochemistry* 56:22-32. 10.1021/acs.biochem.6b00572

Sheils T, Mathias SL, Siramshetty VB, Bocci G, Bologa CG, Yang JJ, Waller A, Southall N, Nguyen DT, and Oprea TI. 2020. How to Illuminate the Druggable Genome Using Pharos. *Curr Protoc Bioinformatics* 69:e92. 10.1002/cpbi.92

Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, Jensen LJ, Vidović D, Koleti A, Schürer SC, Waller A, Yang JJ, Holmes J, Bocci G, Southall N, Dharkar P, Mathé E, Simeonov A, and Oprea TI. 2021. TCRD and Pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Res* 49:D1334-d1346. 10.1093/nar/gkaa993

Simonetti FL, Tornador C, Nabau-Moreto N, Molina-Vila MA, and Marino-Buslje C. 2014. Kin-Driver: a database of driver mutations in protein kinases. *Database (Oxford)* 2014:bau104. 10.1093/database/bau104

Soleymani S, Gravel N, Huang L-C, Yeung W, Bozorgi E, Bendzunas NG, Kochut KJ, and Kannan N. 2022. Dark kinase annotation, mining and visualization using the Protein Kinase Ontology. *BioRxiv*:2022.2002. 2025.482021.

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, and Forbes SA. 2018. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* 47:D941-D947. 10.1093/nar/gky1015

Taylor SS, Shaw AS, Kannan N, and Kornev AP. 2015. Integration of signaling in the kinome: Architecture and regulation of the alphaC Helix. *Biochim Biophys Acta* 1854:1567-1574. 10.1016/j.bbapap.2015.04.007

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney

E, Kohli P, Jumper J, and Hassabis D. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596:590-596. 10.1038/s41586-021-03828-1

U M, Talevich E, Katiyar S, Rasheed K, and Kannan N. 2014. Prediction and prioritization of rare oncogenic mutations in the cancer Kinome using novel features and multiple classifiers. *PLoS Comput Biol* 10:e1003545. 10.1371/journal.pcbi.1003545

Vazquez M, Pons T, Brunak S, Valencia A, and Izarzugaza JM. 2016. wKinMut-2: Identification and Interpretation of Pathogenic Variants in Human Protein Kinases. *Hum Mutat* 37:36-42. 10.1002/humu.22914

Watkins X, Garcia LJ, Pundir S, Martin MJ, and Consortium U. 2017. ProtVista: visualization of protein sequence annotations. *Bioinformatics* 33:2040-2041.

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, t Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, and Mons B. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018. 10.1038/sdata.2016.18

Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, and Wilson M. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 46:D1074-d1082. 10.1093/nar/gkx1037

Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR, Ramaswamy S, Futreal PA, Haber DA, Stratton MR, Benes C, McDermott U, and Garnett MJ. 2013. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 41:D955-961. 10.1093/nar/gks1111

Yonemoto W, Garrod SM, Bell SM, and Taylor SS. 1993. Identification of phosphorylation sites in the recombinant catalytic subunit of cAMP-dependent protein kinase. *J Biol Chem* 268:18626-18632.

Yun C-H, Mengwasser KE, Toms AV, Woo MS, Greulich H, Wong K-K, Meyerson M, and Eck MJ. 2008. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences* 105:2070-2075. 10.1073/pnas.0709662105

Zhang J, Yang PL, and Gray NS. 2009. Targeting cancer with small molecule kinase inhibitors. *Nat Rev Cancer* 9:28-39. nrc2559 [pii]
10.1038/nrc2559

Zhang YC, Huo FC, Wei LL, Gong CC, Pan YJ, Mou J, and Pei DS. 2017. PAK5-mediated phosphorylation and nuclear translocation of NF-κB-p65 promotes breast cancer cell proliferation in vitro and in vivo. *J Exp Clin Cancer Res* 36:146. 10.1186/s13046-017-0610-5

Figure 1

The ProKinO architecture and work-flow.

A) Left panel shows a subset of curated data sources used in ontology population. B) The middle panel shows a schematic of the ontology schema with classes (boxes) and relationships (lines) connecting the classes. C) The right panel shows applications for ontology browsing and navigation.

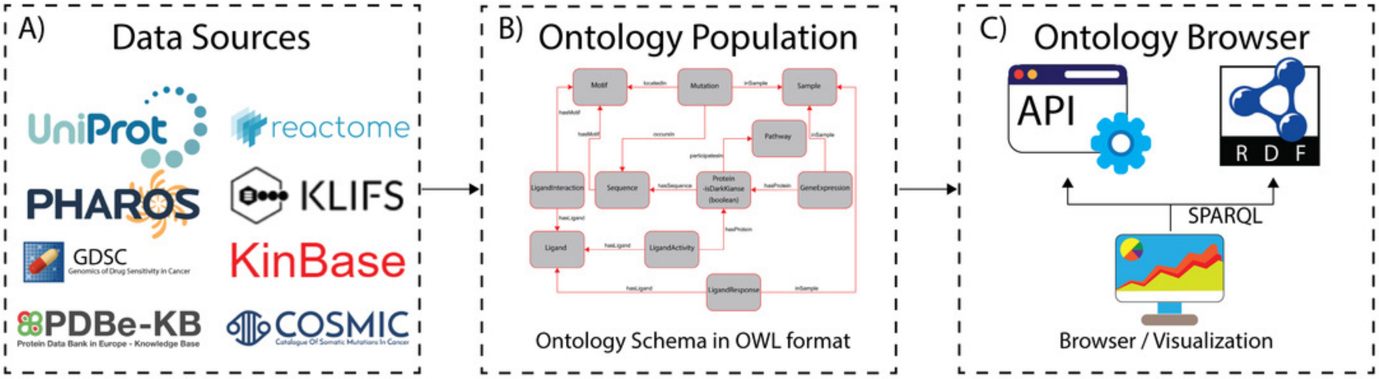


Figure 2

Subset of the updated ProKinO schema with new classes and relationships.

The full schema can be accessed at http://prokino.uga.edu/. New classes are colored in cyan and pre-existing classes are colored in pink. Black arrows indicate new relationships introduced to connect the new classes.

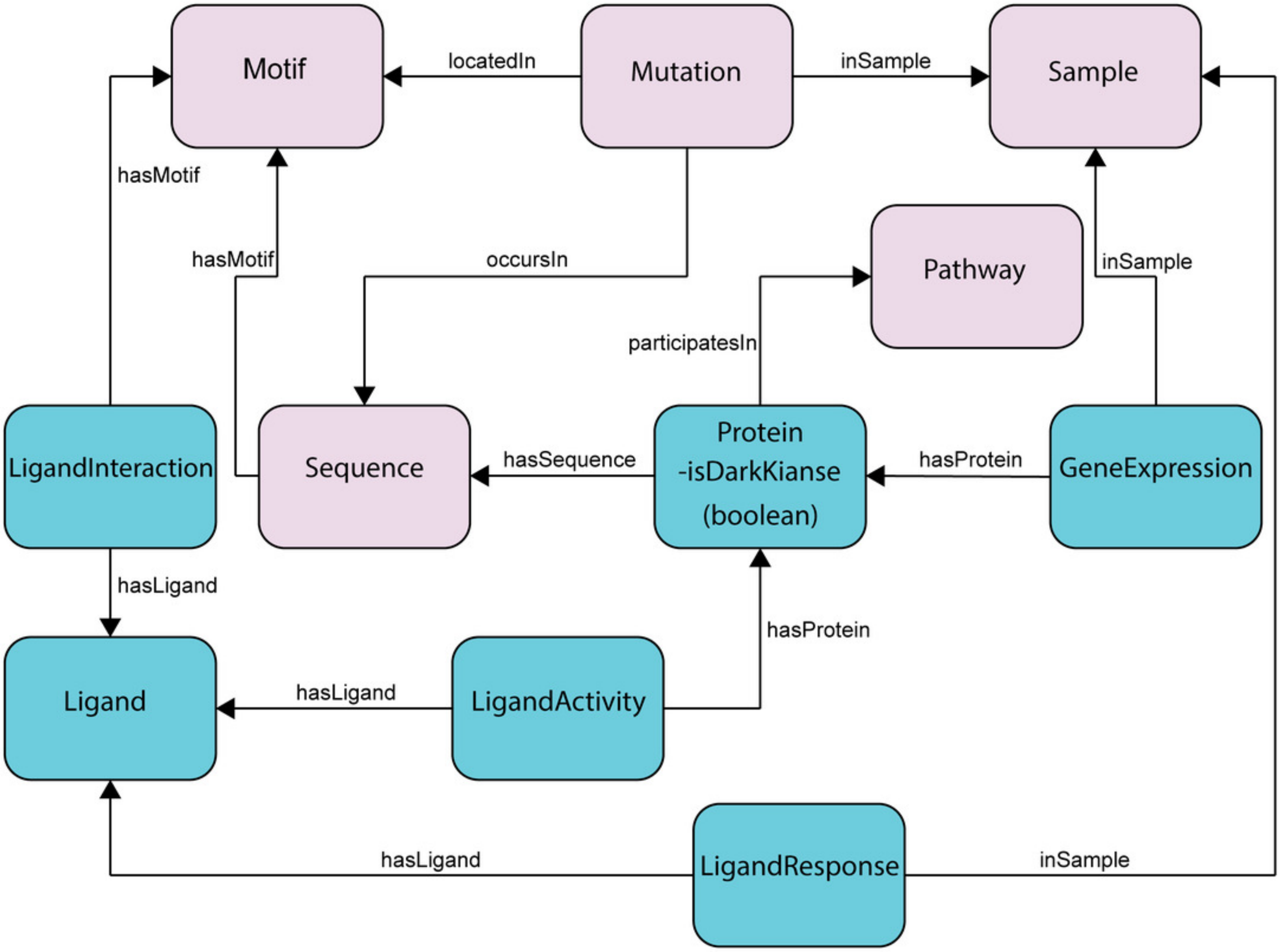
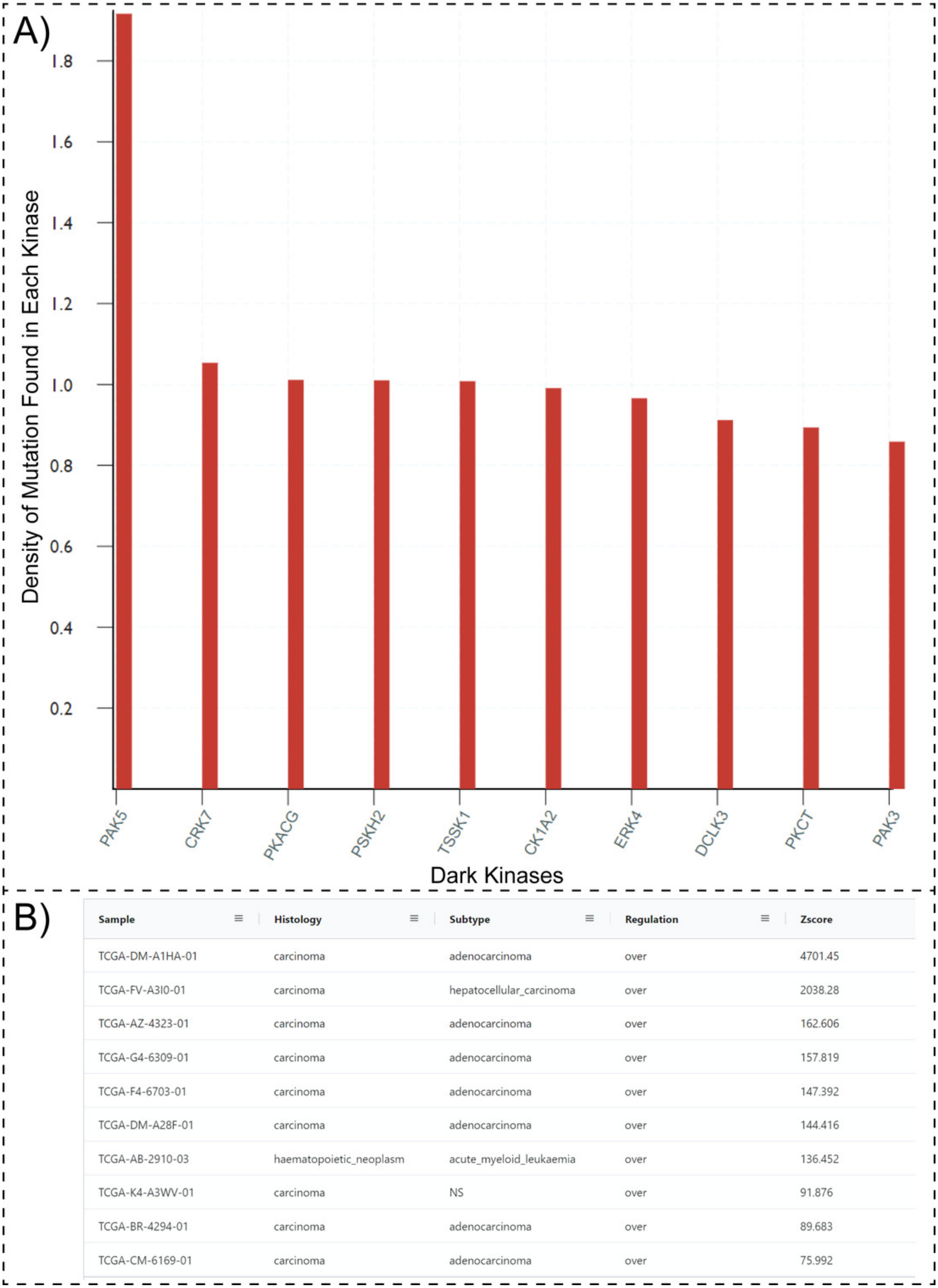


Figure 3

SPARQL query results for Query 27 and 33.

A) Output of Query 27 requesting top 10 dark kinases with most mutations in different cancer types. The mutation counts are normalized by sequence length. B) Output of Query 33 listing samples with abnormal PAK5 expression. The query also lists histology, cancer subtypes, regulation, and Z-score. Only a subset of the query results is shown because of space constraints.

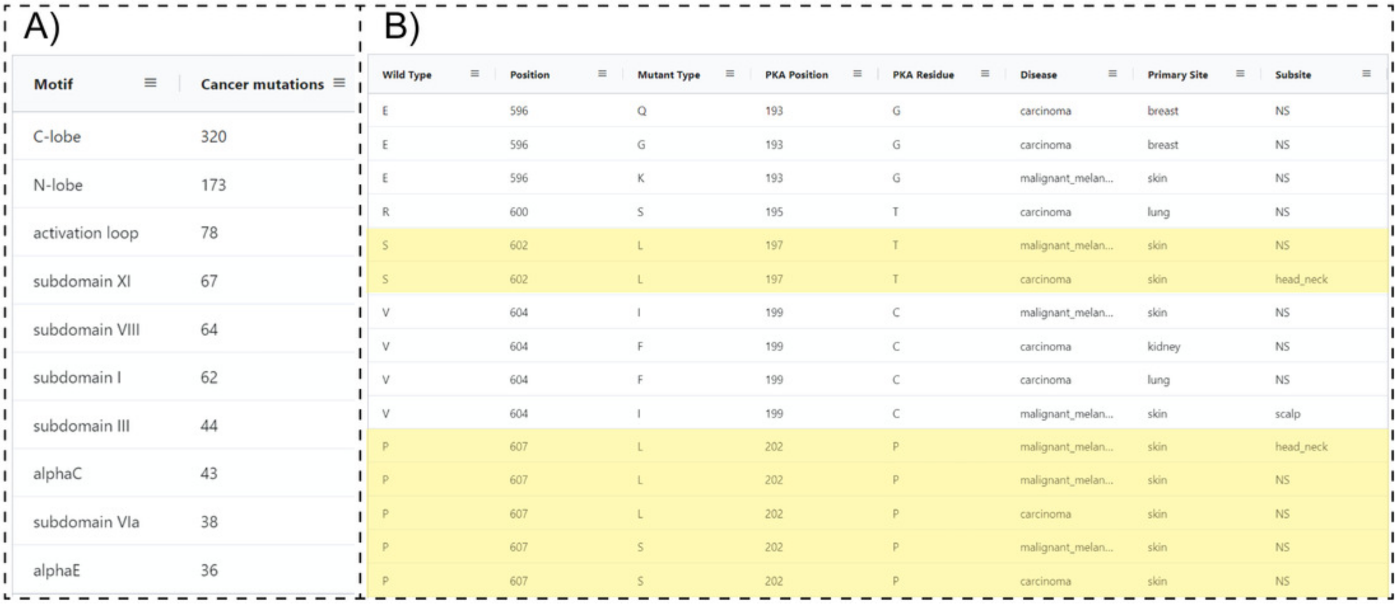


Sample	Histology	Subtype	Regulation	Zscore
TCGA-DM-A1HA-01	carcinoma	adenocarcinoma	over	4701.45
TCGA-FV-A3I0-01	carcinoma	hepatocellular_carcinoma	over	2038.28
TCGA-AZ-4323-01	carcinoma	adenocarcinoma	over	162.606
TCGA-G4-6309-01	carcinoma	adenocarcinoma	over	157.819
TCGA-F4-6703-01	carcinoma	adenocarcinoma	over	147.392
TCGA-DM-A28F-01	carcinoma	adenocarcinoma	over	144.416
TCGA-AB-2910-03	haematopoietic_neoplasm	acute_myeloid_leukaemia	over	136.452
TCGA-K4-A3WV-01	carcinoma	NS	over	91.876
TCGA-BR-4294-01	carcinoma	adenocarcinoma	over	89.683
TCGA-CM-6169-01	carcinoma	adenocarcinoma	over	75.992

Figure 4

SPARQL query results for Query 28 and 29.

A) Output of Query 28 listing the number of unique cancer-linked mutations at various structural locations of PAK5 kinase. B) Output of Query 29 listing unique point mutations in the activation loop of PAK5 kinase. The query also lists the equivalent PKA position, disease type, primary site of the tissue sample, equivalent residue for the PKA positioning of PKA, and subtype of the tissue sample. Entries containing only one mutation per position were filtered from the original query. Only a subset of the query results is shown.



A)

Motif	Cancer mutations
C-lobe	320
N-lobe	173
activation loop	78
subdomain XI	67
subdomain VIII	64
subdomain I	62
subdomain III	44
alphaC	43
subdomain VIa	38
alphaE	36

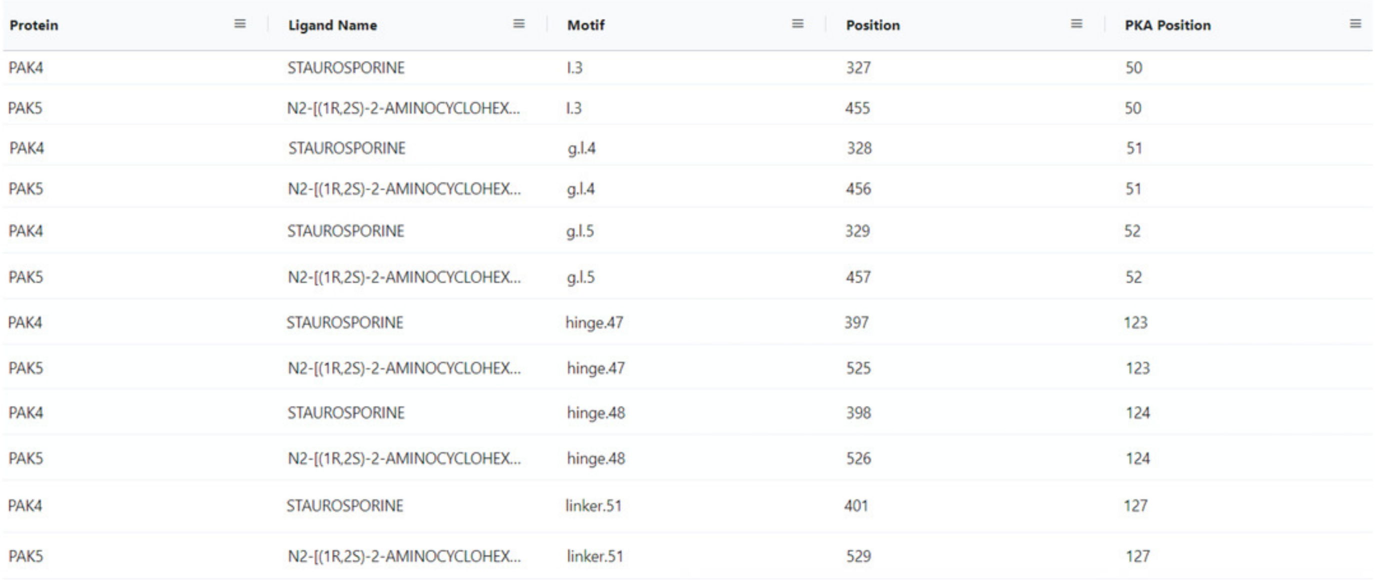
B)

Wild Type	Position	Mutant Type	PKA Position	PKA Residue	Disease	Primary Site	Subsite
E	596	Q	193	G	carcinoma	breast	NS
E	596	G	193	G	carcinoma	breast	NS
E	596	K	193	G	malignant_melan...	skin	NS
R	600	S	195	T	carcinoma	lung	NS
S	602	L	197	T	malignant_melan...	skin	NS
S	602	L	197	T	carcinoma	skin	head_neck
V	604	I	199	C	malignant_melan...	skin	NS
V	604	F	199	C	carcinoma	kidney	NS
V	604	F	199	C	carcinoma	lung	NS
V	604	I	199	C	malignant_melan...	skin	scalp
P	607	L	202	P	malignant_melan...	skin	head_neck
P	607	L	202	P	malignant_melan...	skin	NS
P	607	L	202	P	carcinoma	skin	NS
P	607	S	202	P	malignant_melan...	skin	NS
P	607	S	202	P	carcinoma	skin	NS

Figure 5

SPARQL query results for Query 30.

Output of Query 30 listing ligands interactions with each PAK family member (PAK1-6). It also includes motif names and positions of full sequence and PKA positioning. The output of Query 30 was rearranged to highlight the homology of PAK4 and PAK5 motif/ligand interactions and the figure highlights only a subset of the query results. Run SPARQL query for full results.

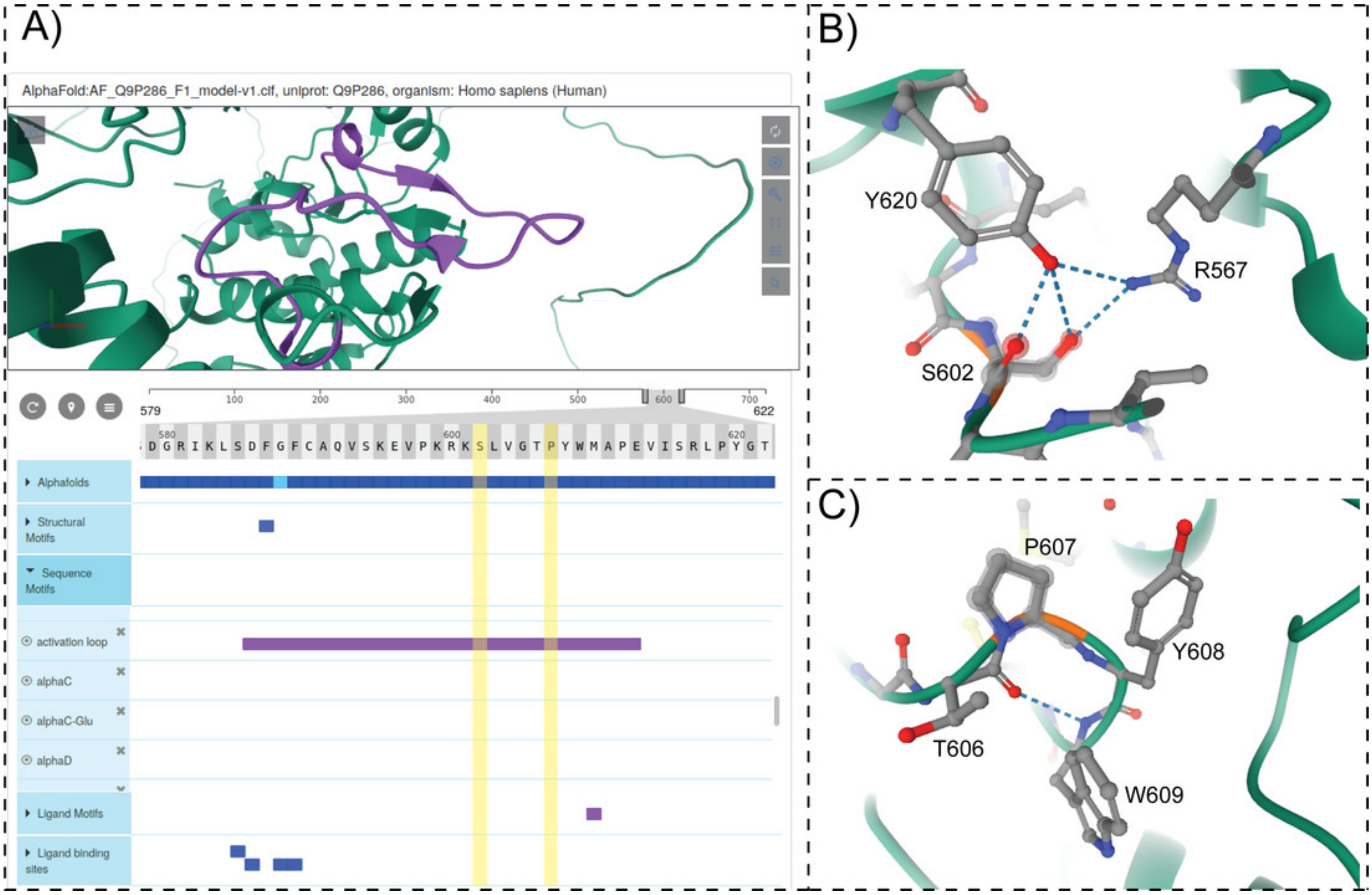


Protein	Ligand Name	Motif	Position	PKA Position
PAK4	STAUROSPORINE	I.3	327	50
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	I.3	455	50
PAK4	STAUROSPORINE	g.l.4	328	51
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	g.l.4	456	51
PAK4	STAUROSPORINE	g.l.5	329	52
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	g.l.5	457	52
PAK4	STAUROSPORINE	hinge.47	397	123
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	hinge.47	525	123
PAK4	STAUROSPORINE	hinge.48	398	124
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	hinge.48	526	124
PAK4	STAUROSPORINE	linker.51	401	127
PAK5	N2-[(1R,2S)-2-AMINOCYCLOHEX...	linker.51	529	127

Figure 6

ProtVista viewer.

A) AlphaFold2 model of PAK5 kinase is shown in the structure viewer (top panel). Sequence viewer with annotations are shown in the bottom panel. B-C) Zoomed in view of structural interactions associated with S602 and P607 in PAK5 activation loop.


A)
AlphaFold:AF_Q9P286_F1_model-v1.cif, uniprot: Q9P286, organism: Homo sapiens (Human)
100
200
300
400
500
600
700
579
622
580
600
620
D G R I K L S D F G F C A Q V S K E V P K R K S L V G T P Y W M A P E V I S R L P Y G T
Alphafolds
Structural Motifs
Sequence Motifs
activation loop
alphaC
alphaC-Glu
alphaD
Ligand Motifs
Ligand binding sites
B)
Y620
R567
S602
C)
P607
Y608
T606
W609