3D structural analysis of RAS effector binding and investigation of context-specific networks

By

Philipp Junk, PhD Candidate

UCD student number: 18206816

The thesis is submitted to University College Dublin in fulfilment of the requirements for the degree of Doctor of Philosophy in Translational Medicine

UCD School of Medicine

Head of School: Professor Michael Keane

Principle Supervisors: Associate Professor Christina Kiel (From 2018 to 2021) and Assistant Professor Melinda Halasz (From late 2021 to 2022)

Members of the Research Studies Panel:
Professor Desmond Higgins
Assistant Professor Dirk Fey

September 2022

Based on research carried out in Systems Biology Ireland, School of Medicine, University College Dublin

Contents

Abstract

RAS is a signalling switch important in tissue homeostasis and cancer. In its active form, numerous effector proteins with a RAS binding domain are able to compete for binding. While some of these effectors such as RAF are well characterized, a systems understanding of the RAS effector system in the context of effector competition and signalling, oncogenic mutations, and co-stimulatory contexts is lacking.

In this thesis, the RAS effector system was investigated from a structural and an experimental perspective. For the structural perspective, 54 putative effector proteins in complex with RAS were modelled. Based on the structural models, binding affinities were estimated which were incorporated into a mathematical model of RAS effector complex formation in 29 different tissues. Interface mutations interfering with the complex structures were energetically evaluated in silico and then integrated in the complex formation modelling.

For the experimental perspective, KRAS interactomes from Caco-2 cells overexpressing different KRAS plasmids (WT, G12C, G12D, and G12V KRAS) and treated with different stimulations/inhibitors (untreated, DMOG, TNF-α, IL-6, PGE2, and EGF) were investigated for functional differences. Then, using random walks on a network of the proteins found in the interactomes, we predicted the contribution of individual effectors to functional processes.

Our work numerically describes the RAS effector system in a methodologically coherent way and lays the foundation to experimentally engineer the RAS effector system with interface mutations to selectively enhance or abolish ("rewire") some effectors. In addition, we describe how different combinations of mutations and treatments shape functional annotations of the KRAS interactome.

Statement of Original Authorship

I hereby certify that the submitted work is my own work, was completed while registered as a candidate for the degree stated on the Title Page, and I have not obtained a degree elsewhere on the basis of the research presented in this submitted work.

Thesis Format

This thesis was prepared as a “Collection of Papers” in accordance with the “Guidelines for Preparation, Submission, Examination and Dissemination of Research Degree Theses”[1] and the “Guidelines on Theses as a Collection of Papers”[2] from University College Dublin.

The guidelines are accessible through the shortened links or through the UCD Document Repository[3].

The thesis consists of a critical overview of the overarching research question, and four manuscripts. On the title pages of each manuscript, detailed information about the authors and where to access it is provided.

The format and length of the thesis has been discussed by the candidate with supervisors and the Doctoral Studies Panel.

[1] https://bit.ly/3KNCQcj
[2] https://bit.ly/3wYJQxi
[3] https://www.ucd.ie/graduatestudies/documentrepository/

Collaborations

The manuscripts in chapters 2 (page 30), 3 (page 47) and 4 (page 58) have been collaborations between me, Philipp Junk, and the principal supervisor, Christina Kiel. For these manuscripts, any data or code generation was performed by me. The conception of the research and the writing of the manuscripts was performed in collaboration.

The manuscript in chapter 5 (page 107) is a collaboration with other members in the lab and institute. I have not been involved in any wet-lab activity, instead I was performing the bioinformatic analysis and method development. Specifically, the following section in the methods part of that manuscript, and their associated results, were my contribution.

- Section 5.2.5 on page 137
- Section 5.2.5 on page 138
- Section 5.2.5 on page 139
- Section 5.2.5 on page 141

The conception of the research and the writing on the manuscript was performed in collaboration (see also Author contributions for this paper on page 142).

Funding

This work is part of the research program “Quantitative and systems analysis of (patho) physiological signalling networks” with project number [16/FRL/3886], which is financed by Science Foundation Ireland (SFI) (to Christina Kiel).

Acknowledgements

I am very grateful for the enormous support I have gotten from my supervisor Christina Kiel over the years. Thanks for allowing me to learn and follow my scientific interests. And for tolerating my treatment of deadlines.

Many thanks as well to Melinda Halasz who was here to step in as my supervisor for the last months.

Many thanks to Des Higgins and Dirk Fey for being a supportive Doctoral Studies Panel.

I want to acknowledge the UCD School of Medicine for funding my PhD.

I am happy and grateful to have become part of such a supportive group, office, and institute. Many thanks to Camille, Cian, Thomas for having been an awesome support! To colleagues past and present, thanks for making science fun: Hugo, Irene, Julie, Kieran, Laura, Leila, Lisa, Luis, Melissa, Miriam, Sarah, Scott, Simona, Solene, Soraya, Steph, Swathi, and everyone else from SBI and Charles.

I want to thank my friends for their support over the years: Ally, Anna, Basti, Dennis, Doro, Franca, Hugh, Jessi, Ju, Julia, Kathi, Katrin, Lari, Lena, Luke, Luna, Martin, Max, Michael, Philipp, Sam, Sam, Stephan, Vero, Xenia.

Thanks for everything, Rita and Gerd!

Finally, I want to honourably mention coffee for making this thesis possible.

List of Publications

Junk P, Kiel C, **in press** at *Structure*. Structure-based prediction of Ras-effector binding affinities and design of branchegetic interface mutations.

Ternet C, **Junk P**, Sevrin T, Catozzi S, Wåhlén E, Heldin J, Oliviero G, Wynne K, Kiel C, 2023. Analysis of context-specific KRAS-effector (sub) complexes in Caco-2 cells. *Life Science Alliance* 6(5)
(CT, PJ and TS contributed equally).

Sevrin T, Strasser L, Ternet C, **Junk P**, Caffarini M, Prins S, D'Arcy C, Catozzi S, Oliviero G, Wynne K, Kiel C, Luthert P, 2023. Whole-Cell Energy Modeling Reveals Quantitative Changes of Predicted Energy Flows in RAS Mutant Cancer Cell Lines. *iScience*, 26(2).

D'Arcy C, Bass O, **Junk P**, Sevrin T, Oliviero G, Wynne K, Halasz M, Kiel C, 2023. Disease–Gene Networks of Skin Pigmentation Disorders and Reconstruction of Protein–Protein Interaction Networks. *Bioengineering*, 10(1):13.

Junk P, Kiel C, 2022. HOMELETTE: a unified interface to homology modelling software. *Bioinformatics*, 38(6), pp.1749–1751.

Junk P, Kiel C, 2021. Engineering of Biological Pathways: Complex Formation and Signal Transduction. *Methods in Molecular Biology*, 2315, pp.59–70.

List of Figures

List of Tables

1 3D structural analysis of RAS effector binding and investigation of context-specific networks

1.1 Introduction

1.1.1 Systems biology

There are many definitions of systems biology.

- “Systems biology is the science that studies how biological function emerges from the interactions between the components of living systems and how these emergent properties enable and constrain the behaviour of those components” attributed to Jan-Hendrik Hofmeyr, in [1].
- “Systems biology means different things to different people. There are those who see it as a logical continuation of functional genomics - that is, carrying out experiments on the genome scale with the aim of understanding how the whole is greater than the sum of its parts. Others see it as a branch of mathematical biology, which consists of the study of small systems for which sufficient parameters have been measured to allow simulations of how the molecules function together to achieve a particular outcome. In our view, it is both of these things.” from [2].
- “Molecular biology has uncovered a multitude of biological facts ... but this alone is not sufficient for interpreting biological systems. Cells, tissues, organs, organisms and ecological webs are systems of components

whose specific interactions have been defined by evolution; thus a system-level understanding should be the prime goal of biology." from [3]

- "For any phenotype - molecular, macroscopic, or ecological - a set of interrelated factors exist that contribute to this phenotype. Since these factors interact, they need to be studied collectively, not merely individually." from [4].
- "... systems biology implies the quantitative understanding of a system, rather than of the individual components, allowing testable predictions to be made. As such, systems biology requires acquisition of data, parameter quantification, bioinformatics analysis and mathematical modelling." from [5].

What most of these definitions have in common, and what is a good consensus for the scope of this thesis, is the separation of individual components of the system, and the behaviour of the system when components of systems interact. Biological systems have been shown to exhibit strikingly complex behaviour, for examples how a whole organism develops from a single cell, or how cells can be reprogrammed as cancer cells and evade both the immune system and therapy, or how our nervous system is able to process stimuli from the outside world and make us react to them. For many of the complex questions in biology, or in other words, for the understanding of many of the complex systems in biology, it is not enough to zoom in on individual components, but necessary to identify core components and understand the behaviour emerging from their interactions.

Models in systems biology

Biological science in its fundamental approach is heavily reliant on model systems, such as cell lines, animal models, or computational models. Models are abstractions of reality, which allow us to investigate relevant aspects of a system while reducing complexity and increasing reproducibility. In the following sections, three different types of computational models for biological systems relevant to this thesis will be introduced.

Protein structures Structural models of biomolecules, in particular proteins, have been useful tools for biology for a long time [6]. There exist different techniques for experimentally determining protein structures, most prominently used are X-ray crystallography and Nuclear Magnetic Resonance (NMR). However, generating these structures takes a huge effort. Structures are deposited in a common database the Protein Data Bank (PDB) [7].

In addition to solving structures experimentally, there are also computational methods for creating structural models of proteins. Historically, it has been more successful to model proteins afters close relatives for which structural information is available, which is called homology modelling [8]. Recently however, there has been a breakthrough with *de novo* structure prediction based on deep learning. AlphaFold2, released in 2021, is able to generate structural models at almost the accuracy of X-ray structures, a development which is currently accelerating and revolutionizing structural biology [9, 10].

Protein structures have been invaluable at explaining the function, mechanism and assembly of proteins or protein complexes at the molecular level. High quality protein structures have been the foundation for structure-based drug design. Finally, knowing the three-dimensional structures of proteins has allowed for the engineering of proteins. There have been applications in the engineering of metabolic proteins with the aim of for example increasing enzymatic activity [11], or the engineering of signalling pathways with the aim of understanding and controlling signalling in a model system [12].

Mechanistic mathematical models Another critical tool for computational biologists and mathematicians trying their luck in the biological sciences (and for everyone else), mechanistic modelling is one of the most powerful ways to understand a biological system [1]. Constructing accurate mechanistic models of complex biological systems requires smartly chosen assumptions, biological domain knowledge and potentially many cycles of experiments and model refinement [1].

Iterative refinement of mechanistic models is a central step in constructing a good model for a problem. Usually, core components of a system would

be identified. Then, their interactions would be described in mathematical notation. Given a set of physical parameters, i.e. the number of molecules of a certain species present, or the reaction rates of a certain reaction between two species, the model could be simulated with the aim of extracting a quantifiable outcome that can be compared to experimental data. The model might be performing well on some subsets of the data, but not on others. Subsequently, the assumptions, species and their interactions need to be re-evaluated in order to identify potential steps for refining the model. Ideally, based on these potential refinements, new experiments can be designed that help verify the correct direction for model refinement. This has been described as the iterative cycle of model refinement, where "experiment-informed modelling" is followed by "modelling-informed experiments" and so on [1].

Successful mechanistic models reward with an understanding of a system and its possible behaviours in detail that is basically unrivalled. Examples for interesting mechanistic models include the dynamic behaviour of signalling pathways [13] or the stratification of patients [14].

Depending on the biological question, different mathematical frameworks of different complexity can be used. For example, modelling time dynamics requires different equations than modelling equilibrium states independent of dynamic changes [1, 13, 15].

Biological networks Networks are collections of nodes and edges that connect the nodes. There are many different types of networks used in biology, such as protein-protein interaction (PPI) networks, or gene regulatory networks [16]. Networks can be constructed in three different ways: 1) based on knowledge from scientific literature (i.e. the STRING database as well-known knowledgebase for PPIs [17]), 2) based on computational predictions (i.e. Kboost for the reconstruction of gene regulatory networks [18]) or 3) based on high-throughput experiments (i.e. the HuRI project based on genome-scale Yeast2Hybrid screens [19]).

Of particular interest to this thesis are PPIs and their reconstruction using affinity-purification mass spectroscopy (AP-MS) [20]. In AP-MS experiments, a

"bait" protein is isolated from a cell lysate together with all proteins binding to it, so called "prey" proteins. The identity of these proteins is then determined by mass spectroscopy. One of the advantages of the technique is that it is applied on cell lysates, making the resulting interaction network specific to the cell type or context. This makes it a great tool for studying context-specific PPIs [21].

The perturbation of interaction network has been linked to human diseases [16]. It was found that there are two different types of perturbations on interaction networks. A) perturbations that removed nodes from the network ("node-removal") and B) perturbations that remove edges from the network ("edgetic") [22]. Edgetic perturbations are linked to proteins that can be involved in multiple diseases, with different mutations removing different edges from the interaction network leading to different functional outcomes.

While challenges remain in the areas of construction and validation of networks, networks have been useful tools for visualizing and analysing how components of a system are interacting.

Combination of different models Of course, these types of models are not necessarily separate from each other but can be complementarily combined to gain better understanding of the biological system under investigation [2]. For example, structural modelling has been shown to predict the pathological changes of point mutations on network architecture in RASopathies and RAS-related cancers [23]. In another example, Rukhlenko et al. used structural modelling to analyse the conformational space of RAF kinases with and without inhibitors, which they could translate into a rule-based modelling approach [24]. Their model was able to accurately describe paradoxical activation of RAF kinase after treatment with RAF inhibitors and suggest treatment strategies to counter it. To conclude, there is much potential in combining different layers of modelling.

The three steps of systems biology

For this thesis, we distinguish three steps in the analysis of the system [4, 25]:

Building of the system At the beginning, for the biological question at hand, an appropriate representation needs to be chosen. As there are many possible models to represent a system with, the main focus should be the fit to the biological questions. Also, the choices made in this step will impose and define the limitations of analysis that can be done in subsequent steps.

Examples for this can be the construction of networks of the system components and their interactions, or the formulation of mathematical equations that describe the interactions of the components. Of course, also wet-lab methods such as the generation of interaction maps from AP-MS are examples of building a system.

Analysis of the system The next step is to analyse the system in order to understand if it comes with the required compromise between abstraction and complexity needed to answer the underlying biological question. Does the system behave in the expected way, or, if data for the biological question is available, is the experimental data captured in the behaviour of the model system?

Perturbation of the system Finally, a big emphasis is put on the perturbation of the system. Choosing the right perturbations can help to validate the model, or to distinguish between competing biological hypothesis [25]. With a well-validated model, exploring perturbations can also be a good way to identify interesting states of the system, especially in cases for which searching for these states with experiments is time-consuming and/or expensive. Ideally, the focus should be on perturbations that have large effects on the system [4].

If necessary, this approach can be performed iteratively, similar to the iterative cycle of mechanistic modelling and experiments.

1.1.2 The RAS system

The biological system under investigation are RAS proteins and their effector proteins, investigated by biological, oncogenic, and synthetic perturbations.

The proteins HRAS, NRAS, KRAS4A and KRAS4B (from the HRAS, NRAS and KRAS genes) are a family of small GTPases and signalling proteins of extraordinary importance. In the following text, these proteins will be collectively referred to as “RAS”.

Functional aspects of the RAS system

RAS proteins are around 190 amino acid (aa) long, membrane-associated, globular proteins [26, 27]. Between the different isoforms, they share an identical effector lobe (aa 1 - 86) and an almost identical allosteric lobe (aa 86-166) (Figure 1.1A). The effector lobe and the allosteric lobe together form the so-called G domain. The N-terminus of the RAS proteins is constituted by the hypervariable region (HVR) (aa 167-188/189). As the name indicates, this region is not well conserved between the isoforms. Functionally, the nucleotide binding region together with the active site is located in the G domain, while the HVR is post-translationally modified to allow RAS the association with/anchorage in the plasma membrane [27].

RAS proteins are molecular switches that, depending on which nucleotide they are bound to, are either active (GTP bound) or inactive (GDP bound) [27]. While RAS proteins are intrinsically able to switch on their own, meaning to hydrolyse GTP to GDP and exchange GDP with GTP, these reactions are quite slow (Figure 1.1B). Instead, there exists a cellular machinery of enzymes to control the state of the RAS switch [27, 28]. GTPase-activating proteins (GAPs) speed up the hydrolysis of GTP, guanine exchange factors (GEFs) accelerate the dissociation of GDP from RAS. Both families of helper proteins increase reaction rates by several orders of magnitude [28].

In the active state, RAS proteins are able to interact with and activate a range of effector proteins [29]. These effector proteins share the same interface (more details in the structural description) and compete with each other for binding on RAS [29, 30, 15]. The most prominent effector proteins are the RAF family of proteins (ARAF, BRAF and RAF1) as well as the PI3K family of proteins. However, besides the well-characterized families, there are actually many more proteins that have been characterized as effectors or potential effectors of

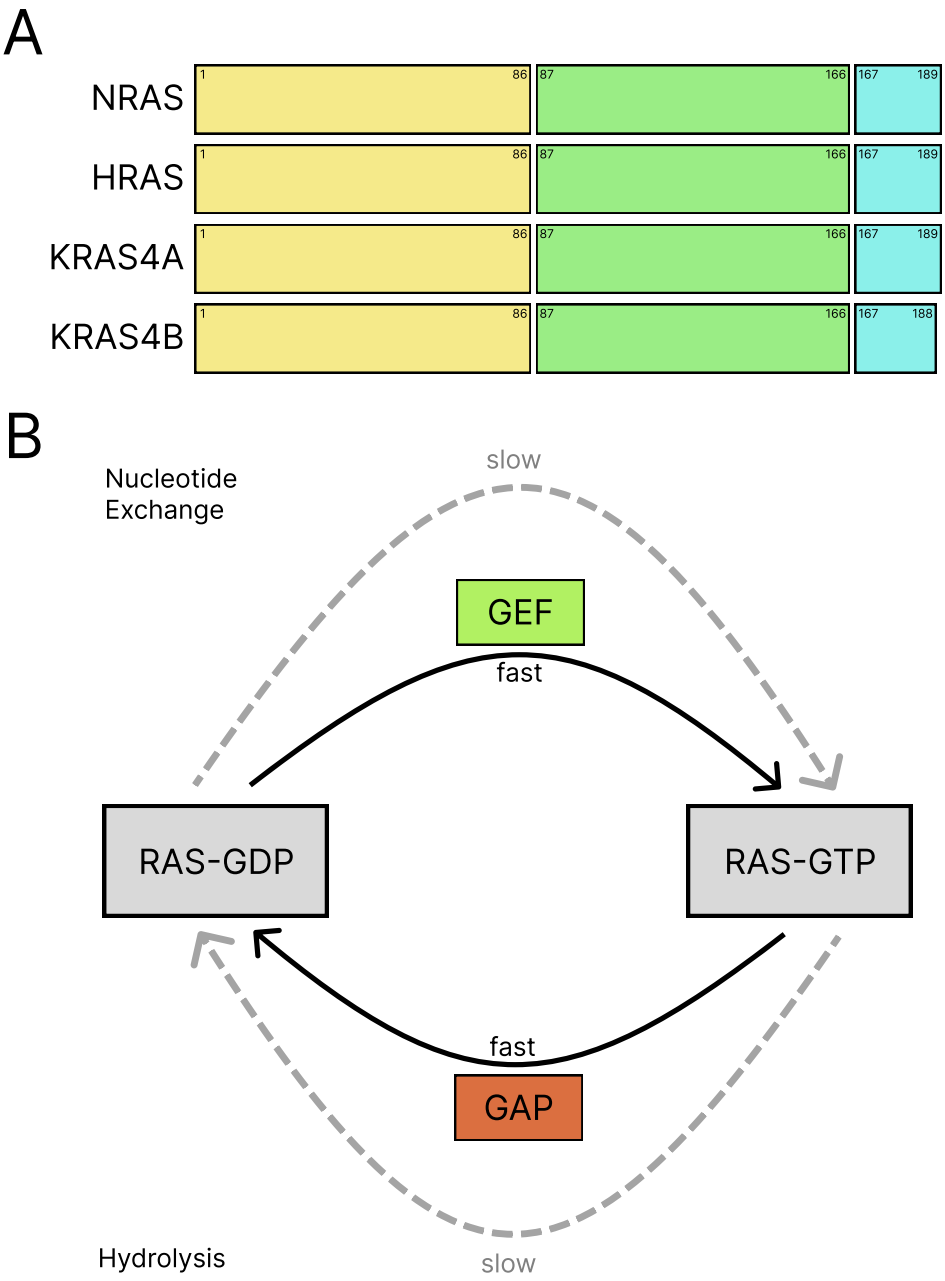

Figure 1.1 Overview over RAS proteins. (A) Separation of RAS in effector lobe (yellow), allosteric lobe (green) and hypervariable region (blue). (B) Overview over the RAS-switch. Intrinsic nucleotide exchange and hydrolysis is slow, but can be sped up significantly with the enzymes GAP or GEF catalysing the reaction.

RAS, such as RALGDS, RASSF5 and AFDN [29, 31, 32, 30]. Through these interactions, RAS proteins are able to influence a large number of cellular functions, including cell proliferation, survival, differentiation, migration and metabolism [29]. Thus, RAS proteins are signalling hubs that play important roles in physiological contexts such as development and tissue homeostasis, but also in pathological contexts such as developmental disorders and cancer [33, 34, 27].

Structural aspects of the RAS system

The structural description of RAS goes back to the late 1980's [35, 36, 37]. Together with an increased structural understanding, many functional aspects of RAS could be explained. For example, the "switching" of RAS has been well characterised structurally [26]. The ability to change the affinity to effector proteins comes from the dynamic assembly and disassembly of the binding interface. These structural changes are mainly facilitated by the switch regions, switch 1 and switch 2, that are differently arranged depending on the bound nucleotide (Figure 1.2). In the GTP-bound (active) state, the two switch regions are tightly associated to the γ-phosphate of GTP by hydrogen bonds. In the GDP-bound state however, these hydrogen bonds are not forming, and the switch regions are looser and further away from the protein core. As most of the binding interface to effector proteins is constituted by switch 1 and 2, only GTP-bound RAS is able to interact with effector proteins [38].

Effector proteins have a domain that enables them to bind to active RAS [31, 38, 29]. There are three families of these domains, namely RAS binding domains (RBDs), RAS association domains (RAs) and PI3K RAS binding domains (PI3K_RBDs) [30]. For the rest of this thesis, they will be collectively referred to as RBDs. RBDs are structurally related to each other, with all of them having the same Ubiquitin-like fold, which is also referred to as Ubiquitin super-fold [40, 38]. These globular domains of size between 70 and 130 amino acids all assume a ββαββαβ-fold [40, 31, 38, 29].

Complexes between RAS and effector proteins seem to be structurally well conserved. As previously mentioned, most of the interface on the side of RAS

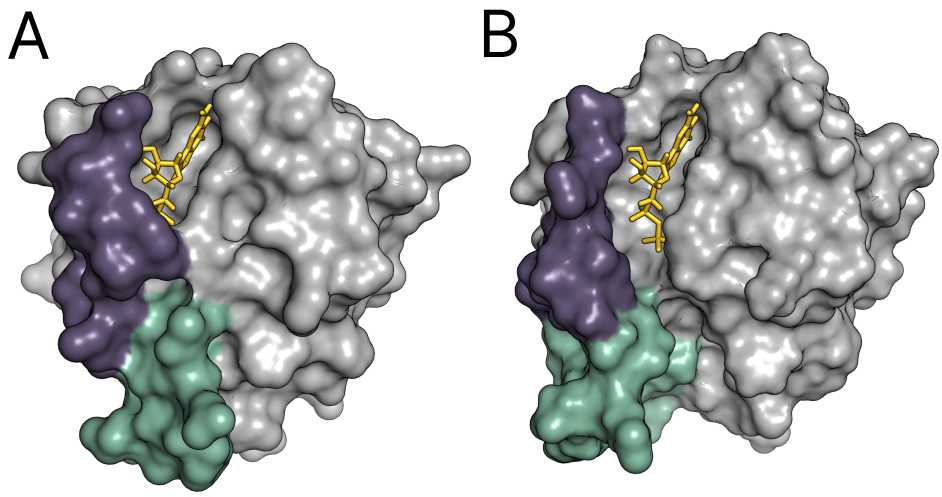

Figure 1.2 Structures of RAS in its active (GTP-bound) state (from PDB: 3GFT) (A) and its inactive (GDP-bound) state (from PDB: 4LPK [39]) (B). RAS protein is coloured in gray, with switch 1 and 2 highlighted in purple and green, respectively. The nucleotide is visualised in gold.

is assembled by switch 1 and 2. The main anchorage point of the interface appears to be the formation of an inter-molecular β-sheet between β2 on RAS and β2 on the RBD. This structural motif is well preserved between all the effector structures that are known [38].

RAS as an oncogene

The RAS protein was originally discovered in the context of causing cancer as a viral oncogene in rats [35]. Later, RAS proteins were identified as oncogenic drivers in human cancer as well. It is now estimated that RAS proteins plays a role in about 25 % of human cancers, with it mostly being found in pancreatic ductal adenocarcinoma (98 %), colorectal carcinoma (52 %), multiple myeloma (43 %), lung adenocarcinoma (32 %) and melanoma (29 %) [27].

As established before, RAS proteins are signalling hubs involved in many different signalling pathways. Dysregulation of this hub and following hyper-activation of downstream signalling subsequently leads to the development of cancer [27]. Dysregulation of RAS occurs by RAS being mutated in one of sev-

eral well-defined oncogenic hotspots. There are three main mutation hotspots which together make up almost the entirety of oncogenic mutations on RAS: G12, G13 and Q61. In simple terms, for all these oncogenic mutations, the mechanistic cause is similar. The cycle of hydrolysis and nucleotide exchange is disrupted, leading to an increase of RAS in the active state, which in turn leads to a hyperactivation of RAS signalling downstream of RAS.

More detailed characterisation however, has revealed that there are numerous differences in the way different oncogenic mutations affect functional parameters. For example, it was shown that for the oncogenic mutations G12V, Q61L and G13D, there are strong differences in their sensitivity to GAP-mediated hydrolysis and GEF-mediated nucleotide exchange [41].

Additionally, changes in signalling behaviour and oncogenic potential have been identified for these and other mutants as well [42, 43, 44, 45, 46]. An overview over differences between the oncogenic mutations can be found in [47]. For example, with regards to biochemical evaluation of effector interactions of RAS oncogenic mutants, it was found that different KRAS mutants displayed slightly different affinities to RAF1, all of them lower than KRAS WT [43]. In a different study, it was found that KRAS G12V mutant had a different hierarchy of which effectors could out-compete each other for binding on KRAS compared to WT [42]. These results indicate that different RAS mutants could preferably interact with some, but not other effectors in a mutation-specific way, an aspect which has not been well explored yet.

On a structural level, none of oncogenic hotspot mutations greatly affect RAS structure, or are present in the interface relevant for effector binding. Instead, oncogenic mutations interfere with the correct assembly of the active site for GTP hydrolysis in RAS-GAP complexes [26]. Additionally, it has been proposed that the dynamics of the RAS structure, in particular the natural movement of the switch regions, is influenced by oncogenic mutations [48]. It is currently however unclear which consequences these impaired dynamics have.

1.1.3 Aims of the thesis

Many aspects of the RAS signalling system still remain a mystery. In particular, our current understanding of the interplay between effector signalling, oncogenic mutations, cellular background and stimulatory context is lacking.

- Which effectors are interacting with RAS?
- In which conditions (with condition being any combination of oncogenic mutation, cellular background and stimulatory context) do these effectors interact?
- Which functional aspects (i.e. proliferation, changes in metabolism, etc.) can be linked to RAS binding to which effectors?

This thesis, based on the associated papers, aims to investigate these questions based on different approaches.

1.2 From structures to systems

1.2.1 Hypothesis and aims

As discussed in the introduction, our understanding of how RAS interacts with its effectors and what can influence this interaction has been lacking. One of the areas that is relatively unexplored is the variety of potential RAS effectors. Some key effectors have been characterized in great detail. However, there are more proteins in the human genome that have a RAS binding domain that might enable them to interact with RAS. For a surprisingly large number of these effectors, we do not know how strongly they interact with RAS. There have been previous efforts to estimate their binding energies to RAS systematically, however, these efforts were not focused on generating protein structures that could be used for complex structural analysis of the whole RAS effector system [49, 50]. We think that this a grey area that should be explored. In particular, we suggest building an *in silico* system on two levels: the first level is the structural level, where for each interaction of RAS with an RBD, structural models will be generated. The second level is the construction of a mathematical model based on parameters from the first layer, that is used to analyse the behaviour of the system as a whole. We hypothesize that building such a two-layer model will enable us to improve our understanding of the RAS effector system. Furthermore, we expect that structure-based perturbations will be a useful tool for identifying interesting ways to engineer the system.

1.2.2 Creating a complete structural representation of RAS-effector interactions

There are certain advantages and disadvantages in using structures for understanding biological systems. At the core of the matter, the structure of a protein or enzyme determines its function. Protein structures are more conserved from evolution than base pair or amino acid sequences are [51]. Therefore, having all structures for a system modelled can provide great insights into the system. Which proteins are interacting? Which enzyme activities are essential for the system? In theory, it should be possible, from protein structures alone

to determine all essential parameters of a biological system: which proteins interact at which rate, which enzyme catalyse which reactions at which speed, how fast is protein folding and degradation, etc. In practice however, it is difficult to get reliable models for all members of a system and even more difficult to describe more complex phenomena such as PPIs. While there are experimental methods for determining protein structures, these have their limitations as well. In addition, deriving functional parameters such as binding affinity between proteins is a non-trivial problem as well. For this reason, structural analysis has been historically more about the investigation of a single protein or complex, and less about the investigation of complex systems. In recent years however, systematic approaches that take 3D protein structures into account have been developed [52, 53].

Our system of interest however is uniquely suited for a systems-level analysis of protein structures and their interactions for multiple reasons. Firstly, the core components of the structural system are relatively simple, being the interaction between RAS in its active (GTP-bound) state with a RAS binding domain (RBD) of an effector protein. Currently, there is a lot of structural analysis on potential modulators of this system, such as RAS interacting with the membrane [54, 55], potential dimers, multimers or nanoclusters of RAS proteins and their effect on stabilizing and destabilizing potential RAS effector complexes [55, 56], recruitment of effectors to the membrane by additional stimuli [57], or additional effector domains interacting with RAS [58]. But, the core component is the pairwise interaction between two protein domains.

Secondly, experimental structures of RAS in complex with the RBD are available for some, but not all potential RAS effectors (Table 4.S1). From these structures, we have learned that PPIs between RAS and its effector proteins interact in very similar ways, resolving around an inter-molecular β-sheet in the interface (Figure 4.1). The orientation of this inter-molecular β-sheet seems to be highly conserved between the different structures, same as the interface on RAS, which is mostly assembled by the functional regions switch 1 and 2. Additionally, although sequence identity between different RBDs is quite low (Figure 4.S1D), the structural fold appears to be well conserved [38, 40].

Thirdly, as a more general point and not for our system specifically, recent advances in the area of protein structure prediction, in particular by AlphaFold2 [9], have proven very useful for our approach. The problem of protein structure prediction has been a major roadblock in structural biology for a long time. While we are able to obtain models from experimental methods such as X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy, each of these come with their own set of caveats. However, the *in silico* prediction of a protein structure seemed for a long time like an unsolvable problem. But, with the release of AlphaFold2 in 2021, the possibility for generating and using high-quality *de novo* predicted structures became a possibility.

Putting these pieces together, we developed a pipeline for the generation of structural models for all RAS RBD interactions (Figure 4.2). The pipeline is based on homology modelling and enhanced by structural predictions from AlphaFold2. A common homology modelling approach requires three types of inputs: 1) Target sequences to model, 2) templates to model them with and 3) a sequence alignment between the target sequences and the templates. I will briefly touch on these.

61 target sequences were used (Table 4.S2). According to a recent analysis by our group, there are 56 putative effector proteins in the human genome. Of these, two do not have an RBD and four have two RBDs, which gives us 58 RBDs to model. In addition, we included three distinct Ubiquitin superfolds from proteins of the Ubiquitin family in order to evaluate how proteins of the same fold, but more distantly related, are modelled in the system.

With regards to the structural templates, we heavily relied on AlphaFold2 in order to enhance the structural predictions. Firstly, we are using the AlphaFold protein structure database [59] to get high quality structures of all of our RBDs. We quality-checked these structures and came to the conclusion that AlphaFold2 is reliable at generating the RBD fold (Figure 4.S1ABC). Secondly, we used AlphaFold2 to generate additional complex templates. While the general orientation of the interface is well conserved, there are minor differences in the way the known complex templates interact with RAS (Figure 4.1BCD). We therefore think it can only benefit our approach if we expand the conformational space to sample from by using AlphaFold2 generated complex

templates that fit in the interaction mode that has been observed so far. After generating complex models for all 58 RBDs using AlphaFold, we picked 36 out of these as additional complex templates for the homology modelling after filtering for alignment of the intermolecular β-sheet (Figure 4.S2). In addition to templates generated by AlphaFold2, we are also using seven RAS effector complex template that were experimentally determined using X-ray crystallography (Table 4.S1).

Lastly, sequence alignments were generated dynamically between each pair of single template structure and complex template structure using TMalign. With these inputs, we used MODELLER, in particular an extension called altMOD which slightly modifies the MODELLER object function, to model our complex structures of interest. For each combination of single and complex template, 300 models were generated. After model generation, a two-step procedure was implemented to select high-quality, representative structures from the thousands of models generated. In the first step, model quality is assessed by several metrics on a structural and stochastic level. For the second step, the binding interfaces for the remaining models are energetically characterized by using FoldX to calculate *in silico* binding energies and alanine scan profiles, which are then processed by an unsupervised learning pipeline. For one of the targets, RASSF3, we came to the conclusion not to trust our structural predictions and removed the structure from further analysis. At the end, for 60 of our targets, we selected three representative structures.

1.2.3 Structural characterization of the system

Following model generation, the next objective was to structurally characterize the complex models. Firstly, we used FoldX to determine energetic hotspots in the interface, in particular in the RAS interface (Figure 4.3AB). We were able to identify 12 different energetic hotspots in at least one of the 60 targets, the well-characterized energetic hotspots I36, Y40 and D38 found the most frequently. Interestingly, I36 and Y40 are exclusively energetically favourable for the formation of the RAS effector complex, whereas D38 can also be unfavourable in a substantial number of structures.

Based on the energetic contributions of residues in the interface, we used a supervised learning regressor to predict binding affinities. Thanks to biochemical measurements of binding affinity between RAS and its effectors over the last 20 years, we were able to collect a set of 22 known binding affinities which we used as a test and training set (Table 4.S5). Our final regressor has a cross-validation R2 score of 0.53 in the training set and an R2 score of 0.77 in the test set. For more details on the machine learning, please refer to section 4.2.6 (on page 84). Using the trained regressor, we were able to predict binding affinities for the full set of 61 RBDs (Figure 4.4, Table 4.S5). All binding affinities are in a reasonable range. There are some surprises, as for example RASIP1 has been previously reported as a non-binder in an experimental measurement but is predicted with a binding affinity of 2.65 µM. This analysis however is the first to assess the whole set of binding affinities for the RAS effector system based on a coherent framework.

Next, we investigated the energetic contributions of the switch regions to the binding of different effectors. This is of interest to us for two reasons: 1) We already observed in the X-ray structures that these contributions can be diverse; and 2) it has been reported that oncogenic mutations influence the dynamics of the switch regions [48] so this could be interesting in the context of why different oncogenic mutations might interact with effectors in a mutation-specific way. We found that the contributions of the switch regions are diverse in our models. Interestingly, when combining this data with the predicted affinities, we observe that good switch 1 energies seem to be associated with low K_d values, whereas this does not hold true for switch 2 contributions (Figure 4.5).

Finally, we were interested in exploring mutations on RAS that could influence the binding with effectors. The rationale for this is to use protein engineering to stir the RAS effector system in interesting directions. For example, what effect would a RAS protein have on an *in vitro* system if it could not interact any more with its usual effectors. We have previously proposed framework and methodology for the *in silico* identification of these branch pruning mutations (see 2 on page 30). Briefly, we used FoldX's "Build model" function to mutate interface residues to all other residues and assess changes in binding energy for all complexes. Then, we removed mutations that were evaluated by FoldX

to impair protein stability. Following this protocol, we were able to identify 200 interface mutations that affected the binding energy with at least one of the 60 complexes (Figure 4.6). Interestingly, we were able to identify mutations that decrease the binding affinity to some effectors and increase it to others (Figure 4.S8B). Analogous to what we observed with the interface hotspot mutations, mutations in I36 almost exclusively decrease the binding affinity to effectors. Y40 mutations are not present that much because many of the Y40 mutants are predicted to impair RAS stability. D38 mutations are very diverse in their effect, both increasing and decreasing binding energy.

1.2.4 Integration of structural parameters into mechanistic modelling

Moving on to the second layer of our description of the RAS effector system: a mechanistic model describing which effectors are bound to RAS in equilibrium in specific biological contexts. There are two components needed for such a model, binding affinity and protein abundances. The model is adapted from previous work in our lab ([30, 15]), this is also where we obtained the abundances for RAS and effector proteins in 29 different tissues from [60]. For the affinities, the predicted affinities based on the structural models determined here were taken. Also, since only GTP-bound, active RAS is able to interact with effectors, we need to define the amount of RAS that is GTP loaded. For this study, we considered two different loads, 20 % and 90 % for a system without and with oncogenic mutation, respectively. For more details on the mathematical description of the system, please refer to 4.2.6 on page 86.

First, we simulated the normal state of the system with our parameters. We find that mostly RAF proteins with their combination of low binding affinity and high protein abundance dominate the binding to RAS across all tissues, which is in line with what has been reported before (Figure 4.7A). Next, we simulated the system with the introduction of an interface mutation, in this example D38A, which has been reported to abolish RAF binding in an experimental context before, which is in line with our *in silico* analysis. For this system, as expected, we see that binding to RAF proteins is almost not observable (Figure 4.7B).

Instead, other effectors such as PI3KCA, RALGDS, RASIP1, RASSF5, RGL2 or SNX27 play more prominent roles in a tissue-dependent manner.

Performing this analysis for all combinations of our 200 interface mutations in 29 tissues and 2 RAS-GTP loads results in many directions the RAS effector system could possibly be rewired to. In order to assess which changes are possible for individual proteins, we calculate for all combinations the difference in bound protein for each effector compared to the system without an interface mutation (Figure 4.S10). We find that for almost half (26 out of 54) of the effectors, it is possible to increase the percentage of protein bound to RAS by more than 25 % of total RAS in the system, and for 9 out of 54 effectors, it is possible to increase it by 50 % or more.

Finally, we visualized all solvable system on a 2D plane with UMAP and organised them in outliers and clusters using OPTICS. We were able to identify 19 cluster (Figure 4.7C) with our settings. These clusters correspond to distinct states the system can occupy in response to our structure-based perturbations. We further explored two of the clusters and showed that these different states can be assembled in different ways. For the cluster analysed in Figure 4.7D, we conclude that its systems come from a diverse tissue background but share similar mutations in the D38 energetic hotspot. The cluster analysed in Figure 4.7E on the other hand is derived from only one tissue (lymph node) with a diverse set of interface mutations.

1.2.5 Summary

Following our objective for this part of our analysis of the RAS effector system, we built the proposed two-layer model, combining structural models with mathematical modelling. We integrated state-of-the-art protein structure prediction software AlphaFold2 into our homology modelling pipeline. We characterized these structural models and used them to estimate binding affinities. The binding affinities were combined with protein abundances to construct a mathematical model describing the amount of steady-state RAS effector complexes in 29 different tissues. Then, we integrated the effects of interface mutations on RAS, which were evaluated on the structural side of the model, into the mathe-

matical model to assess the potential for rewiring the system. The combination of structural and mathematical modelling enabled us to analyse the system in a much deeper way then either one of the layers alone would have allowed us.

The two-fold modelling approach came with some advantages and disadvantages. One of the advantages is that any perturbation to the system that can be modelled on the structural level can be integrated into the model. This includes surface mutations as shown here but could also encompass more complex perturbations such as drugs or maybe the presence or absence of other proteins.

One of the major disadvantages is that this modelling approach is based purely on *in silico* analysis. We have tried to mitigate this by incorporating validations where possible. However, no specific validation has been performed for these finding, nor is it directly integrating experimental data. We are hoping in the future to assess some of the prediction we make in *in vitro* experiments.

In contrast, or rather as alternative to this, the next chapter will introduce a different approach to model the RAS effector system, one that is built upon experimental data.

1.3 From experiments to systems

1.3.1 Hypothesis and aims

For the next step of this analysis of the RAS effector system, we will leave the comfortably deterministic nature of *in silico* modelling behind and investigate the KRAS effector system in an *in vitro* cell line model. Specifically, this chapter is about the analysis of the interactome of KRAS in a Caco-2 cell line model, under different conditions.

While our question from the last chapter, which effectors are interacting with RAS, is still of interest here, there are other questions that need to be asked in the broader context of this system. Moving on to an *in vitro* system makes it possible to investigate the functional consequences that arise from the interaction of RAS with different effectors, but also from other influences. As we pointed out in the introduction, we believe that the outcomes of RAS signalling arise from a complex interplay of effectors, oncogenic mutations, cellular context and external stimuli. We hypothesize that, by experimentally determining the interactome of different RAS mutants under different stimuli using AP-MS, we will be able to shed some light on this interplay. Additionally, we aim to find out which effectors of KRAS are mediating a certain cellular function/biological process/phenotype in a specific context.

1.3.2 The experimental system

In order to understand the functional analysis of our experimental system, in a first step we need to describe the system and the limitations inherent to it.

The analysis is based on the Caco-2 colorectal cancer cell line. Caco-2 cells represent a cell line at the early stage of oncogenic transformation, being mutated in APC, CTNNB1 (β-catenin), SMAD4, and TP53, but not in KRAS [61, 62, 63, 64]. In our system, there are two major factors that we are analysing: the first one is the mutation status of KRAS. As already discussed, oncogenic mutations lock KRAS in its active state and thereby consistently activate downstream signalling pathways. We also discussed that there seem to be functional differences between the different oncogenic mutations, not just

on the level of different mutated residues (i.e., not just G12 and G13 mutations are different, but also G12V and G12D for example). In order to investigate this influence on the KRAS signalling system, the following plasmids were used that enable exogenous expression of KRAS in Caco-2 cells: flag-KRAS-WT, flag-KRAS-G12C, flag-KRAS-G12D and flag-KRAS-G12V. We also discussed the potential influence of co-stimulatory signals on effector recruitment and therefore downstream signalling of KRAS. In order to model this influence on the system, the cells were starved pre-transfection and then exposed to specific treatments for 24 h until harvested. The following treatments were applied: Dimethyloxalylglycine (DMOG), Interleukin-6 (IL-6), tumour necrosis factor-alpha (TNF-α), epidermal growth factor (EGF) and prostaglandin E2 (PGE2). These treatments were accompanied by an untreated control. DMOG is a drug that inhibits PHD1-3, which leads to the stabilization of HIF-α and thereby partially mimics the cellular hypoxia response [65, 66]. IL-6 and TNF-α as cytokines as well as PGE2 as a prostaglandin are common modulators of the immune system in the colon. Finally, EGF is an important growth factor regulating epithelial homeostasis, which is often de-regulated in an oncogenic context. All these treatments were selected to mimic relevant conditions in the (patho-) physiological environment. See [34] for an extensive review on the roles and influences of the selected treatments on homeostasis and/or oncogenic transformation in colorectal tissue.

The experimental objective was the identification of the KRAS interactome in the above-described conditions. In brief, the following protocol was used. For more details, please refer to the methods section on page 131. Cells were transfected and treated for 24 h, then harvested and lysed. Immunoprecipitation with the flag-tag was performed. Afterwards, samples were prepared for mass spectroscopy analysis using a standard protocol. After mass spectroscopy, peptides were assigned using MaxQuant, and label-free quantification (LFQ) was perform using MaxLFQ [67, 68]. Finally, the data was filtered and pre-processed using R, DEP and limma [69, 70, 71, 72].

To clearly point out my contributions to this project, I have not been involved with any wet lab work. I was responsible for the bioinformatic and computa-

tional analyses downstream of data collection. A detailed description of my contribution to the paper can be found in the preamble on page vii.

1.3.3 Functional analysis

The first objective of the bioinformatic analysis of this data set is to understand from the interactome data which functional outcomes are influenced. In order to do this, two approaches were chosen: the first approach is a differential interaction analysis, comparing two of our 44 conditions at a time, and then performing a gene set enrichment analysis (GSEA) for each of the comparisons. The second analysis is to sum up the measured LFQ intensities on different functional ontology terms, and then statistically test for differences between the conditions. The detailed procedure for this analysis is described on page 138, see also figure 5.4A for an overview over the analysis.

Both analyses come with their own set of advantages and disadvantages. One drawback of the differential analysis is that it is based entirely around binary contrasts, which makes it difficult to capture bigger trends in the data set. On the other hand, it is probably more suited to detect orchestrated changes in individual functional processes. The advantage of the second approach is that due to the statistical procedure, all influences can be considered at once. However, it is not necessarily clear if summed up LFQ intensities are a valid proxy for activity of a certain functional process. Changes in summed up intensities come mostly from absence or presence of individual proteins from the interactome, which does not have to coincide with lower or higher activity in a biological process. Both approaches applied to the whole data set yield many significantly influenced functional processes (Figure 5.S8A).

Next, both analyses were combined by using GO semantic analysis. The idea of semantic analysis for GO terms is to measure the similarity between different GO terms, and to group similar ones together to show combined trends in the data set (Figure 5.4B). Based on semantic similarity, we clustered the 2135 GO terms for which we identified significantly differences into 12 functional clusters. The main five clusters are about metabolism (cluster 1), signal transduction and cellular responses (cluster 2), intracellular transport (cluster 3), differentiation

and development (cluster 4) and cytoskeleton organisation (cluster 5), which cover 1802 of a total of 2135 GO terms (Figure 5.S8D-H).

There are three different factors influencing the interactome of KRAS in our experimental set-up: 1) the mutation status of KRAS, 2) the selected treatment and 3) the concentration at which the treatment was applied. In multiple stages of the analysis, similar conclusions about these factors can be drawn: A) The biggest differences in the mutation status are against the WT, between the oncogenic mutations, these differences are less pronounced; and B) the main influence on the interactome of KRAS comes from the treatments and is enhanced by the overexpression of any oncogenic mutation. These trends are in line with what can be observed on the PCA of the LFQ intensities: the variability for the WT is relatively small, whereas across the oncogenic mutants the treatments have comparable positions in the PCA plot (Figure 5.1D). This is also supported by the analysis of which RAS effectors appear in which conditions (Figure 5.3). Finally, these conclusions are supported by the results from the statistical analysis. For both approaches, mutant versus WT and untreated versus treated showed the highest number of significantly different processes (Figure 5.S8B-C).

In order to make the analysis accessible, an interactive R shiny app was developed (Figure 5.S9). Users can explore the data and search for functional processes that might be interesting to their research. The semantic similarity heatmap is also available in an interactive fashion. After selecting a process, the app shows information about the LFQ intensities of the proteins associated with this functional process, as well as the results of the statistical analysis. All of this data can be conveniently exported from the app.

After exploring the data together with my colleagues, we selected some functional processes involved with well-established hallmarks of cancer cells and for which cell assays exist for validation of the functional predictions in the wet lab. In particular, we were interested in cell proliferation and metabolic changes such as glucose metabolism and ATP generation. The wet lab validation was performed by Thomas Sevrin. It was shown that the functional features we tested were in line with our expectations based on the functional analysis of the KRAS interaction (Figure 5.5). This established our data set as a useful tool to

probe for functional features that are influenced by the oncogenic mutations and treatments we have been exploring.

1.3.4 Mapping information flow

Having established that there are phenotypical differences in the experimental system that can be predicted from the functional analysis, the next objective is to describe how these differences are mediated by KRAS. Since we are influencing the system by introducing KRAS while exposing the cells to selected treatments, and we are measuring the interactome of KRAS, we would like to map out how KRAS mediates these connections in its interactome to other proteins that can be predictive of changes in the phenotype. In particular, we are interested in understanding which effector proteins are involved in a particular process.

To this means, we developed a methodology mapping the information flow between two nodes in a network by using biased random walks. There is already a lot of information on downstream signalling of RAS in various databases and network references. For this work, we were using the STRING database filtered for a specific confidence cut-off. Our random walks were always starting from KRAS and going to the other proteins that were found in the AP-MS. We biased the system by favouring nodes in the random selection that were found in the interactome of KRAS in a specific sample. This was done in order to guide the random walks along paths of interacting proteins that are found in the data set. The details of the implementation can be found on page 141.

This method was applied to the ontology terms that were selected for validation from the functional analysis. As a result, we were able to generate context-specific networks of how KRAS interacts with the proteins of interest that are involved in a specific functional process (see Figure 5.S12, 5.S13, 5.S14). Comparing these networks across different conditions, we are able to identify differences in which functional processes are associated with which effectors in which conditions (Figure 5.6). For example, we find that, for the regulation of glucose metabolism, ARAF and RAF1 are important nodes in the networks of the KRAS mutants. However, for the regulation of proliferation,

AFDN seems to play a bigger role than RAF proteins. Interestingly, for the regulation of proliferation in the DMOG-treated samples, PIK3CA is heavily traversed. In conclusion, we are able to create an estimation of how information flows from KRAS to the proteins involved in a functional process and which effector proteins are involved, for the overexpression of different variants of KRAS and the exposure to different treatments.

1.3.5 Summary

Regarding the first objective of this analysis, linking changes in the KRAS interactome in different contexts (induced by overexpression of oncogenic KRAS and selected treatments) to changes in functional outcomes, we employed two different approaches (ANOVA based and differential-binding based) and then combined them using semantic analysis. Overall, our analysis supports the hypothesis that different co-stimulatory signals can have strong effects on the interactome of KRAS. For our analysis, the effect of treatments is bigger than the effect between different oncogenic mutants. The results of this analysis were made available in an interactive R shiny app. Then, in order to understand how KRAS is involved in these functional changes, we mapped the information flow between KRAS and proteins of interest and investigated which effectors were involved.

In conclusion, our analysis improves our understanding of how KRAS is able to induce functional changes through its downstream signalling in different (patho-) physiological conditions in a relevant *in vitro* system.

1.4 Discussion

As already mentioned in the respective summaries, the two analyses address different aspects of the questions we raised in the introduction. While the structural approach to the RAS effector system investigates the RAS-effector interactions as well as the potential to “structurally” perturb them in great detail, the experimental approach explores the interplay between different influences on the RAS signalling and how they are affecting functional outcomes.

Both approaches fit into the systems biology framework of 1) Building the system, 2) Analysing the system, and 3) Inducing perturbations to the system, as described in the introduction.

For the structural analysis, an existing network of PPIs around RAS was extended by proteins with related domains (RBDs) in order to build a RAS-centred PPI network. In a next step, this PPI was structurally modelled using homology modelling. As the final step in the model building, a mathematical model to solve the equilibrium state concentration of effector binding to RAS based on binding affinities and protein abundances was formulated. For the analysis of the system, we characterised the binding interface in the structural models, which we used to estimate the binding affinities for the pairwise RAS-effector interactions using machine learning. Combining the estimated affinities with protein abundances for different tissues from high-quality proteomics [60] enabled us to numerically solve the steady-state equations and quantitatively describe the system in terms of RAS-effector complexes. Finally, we investigated “structure-based” perturbations in our system. Different interface mutations on RAS were evaluated for all complexes, and the respective changes to binding energy with RAS were applied to the mathematical model. We were able to show that these structural perturbations are able to strongly rewire the RAS-effector system depending on the tissue and the respective mutation.

The experimental analysis is organized according to our systems biology framework as well, although the different parts are arguably more interconnected. From a relevant cell line model, the Caco-2 cell, AP-MS was performed for the KRAS-bait protein for a series of interesting perturbations, namely the

different combinations of selected treatments and overexpression of oncogenic KRAS. We investigated which effector proteins could be found in the different conditions. Next, functional analysis was performed in order to evaluate changes by the perturbations in relation to the control conditions. Finally, after integrating another network resource, the STRING [17], into the interactome data, information flow between RAS and functionally interesting proteins was evaluated and compared against the unperturbed system.

Biochemical and functional studies have indicated that different RAS oncogenic mutations display a “mutation-specific” rewiring by modulating binding affinity in a mutation-specific way [43] or changing the outcome of effector competition for binding on RAS over each others [42]. As oncogenic hotspots on RAS are not part of the interface to effector proteins, the most likely explanation remaining is that mutation-specific changes in the dynamics of the RAS protein play a role [48]. These changes, which ultimately result in the population of different conformational states or altered transition rates between these states, could lead to the observed modulation of binding affinities to effectors. Having a full set of RAS effector complex structures, it could now become feasible to investigate this phenomenon. However, the effect sizes observed so far in the only setting where actual binding affinities were measured were probably on the edge of what could be detected with current methodology (binding affinity for RAF1 for KRAS-WT: 0.056 µM, binding affinities for RAF1 for various KRAS-mutants: 0.067 to 0.348 µM [43]). Additionally, as this a problem related to the dynamics of RAS structures, analysis of static proteins structures will not be sufficient and running long molecular dynamics simulations for many effector complexes and many mutants is computationally expensive. Nonetheless, this seems to be another way RAS proteins and their oncogenic mutants are able to modulate for which the set of RAS effector complex structures generated could prove useful.

Over the recent years, an increasing number of structural “modulators” of the RAS effector interactions has been brought to light. These are, 1) the formation of dimers or nanoclusters in the membrane have been reported to stabilize RAS effector complexes [73]; 2) the orientation of RAS to the plasma membrane [54, 55]; 3) the inclusion of more domains then the RBD in the interaction with RAS,

as has been observed with the CRD domain of RAF1 [58]. We propose that, if structural models exist, the RAS-effector complex structures generated in this work could be useful for the integration in these models. For example, recently a model for a helical assembly of RAS proteins in complex with RAF1 coined signalosome was proposed [56]. While the authors focus on the RAF family of effectors and acknowledge that their structural model could not accommodate proteins of the PIK3 family, it would be interesting to evaluate if other effectors could take part in this assembly.

To conclude, with the recent advances in structural biology [9], we expect more systems to be fully modelled on the structural level. We propose that the framework for “branchegetic” rewiring can prove to be a useful tool in other system relying on PPIs.

2 Engineering of Biological Pathways: Complex Formation and Signal Transduction

2.1 Introduction

Signals in signalling networks are usually propagated by the interactions (and potentially subsequent modification) of proteins. When investigating signalling networks, synthetic biology approaches can be useful tools [12]. For example, it can be of interest to manipulate the network in a way that only selective interactions are lost. One approach to this is to use interface mutations, that only affect the complex formation to some, but not all interaction partners. Here, we call these mutations “network-rewiring” mutations.

The following paper describes the methodology of designing “network-rewiring” interface mutations *in silico* using FoldX. As a practical example, three RAS effector complex structures deposited in the PDB are analysed for branch pruning mutations.

2.2 Engineering of Biological Pathways: Complex Formation and Signal Transduction

Philipp Junk[1], Christina Kiel[1]

[1] Systems Biology Ireland and UCD Charles Institute of Dermatology, School of Medicine, University College Dublin, Dublin 4, Ireland

---

The Manuscript was **published** in **Methods in Molecular Biology: Computational Design of Membrane Proteins** on the 25/07/2021.
It is available at `https://doi.org/10.1007/978-1-0716-1468-6_4` under an Open-Access license.

The content is identical to the version published, with changes to formatting and minor changes to bibliography. The supplementary scripts were not included in the thesis.

2.2.1 Abstract

The rational in silico design of interface mutations within protein complexes is a synthetic biology tool that enables—when introduced into biological systems—the artificial rewiring of biological pathways. Here we describe the three-dimensional structure-based design of "rewiring" mutations using the FoldX force field. Specifically, we provide the protocol for the design and selection of interface mutations in three Ras-effector complex structures (PDB entries 3KUD, 4K81, and 6AMB). Ras mutations that impair binding to some but not all interacting partners are selected.

2.2.2 Introduction

Protein design is a valuable molecular tool in synthetic biology, which can be either used to study how natural signaling networks function or to build synthetic networks with specific functionalities [12]. Mutations impacting complex formations can be used to alter the affinities and specificities of an interaction, thereby inducing network "rewiring." Provided that high-resolution three-dimensional (3D) structures of protein complexes are available, the impact of mutations can be quickly and inexpensively assessed using in silico design algorithms such as FoldX [74, 75]. Protein interface modelling using FoldX has been used previously to predict the impact of interface mutations (Ala scan) with good correlations between in silico prediction and in vitro experiments [76]. Likewise, interface mutations with changed electrostatic properties and altered kinetics of complex formations were designed using FoldX and subsequently engineered and tested in human cell lines [77] and yeast cells [78].

Here, we provide the protocol for designing rewiring mutations using FoldX for three Ras-effector complexes, with the aim to disrupt binding of one of the effectors to Ras but keep binding of the other two effectors unchanged for all three effectors (Fig. 2.1). The workflow contains five steps: (1) preparation of 3D complex structures, (2) identification of interface residues, (3) evaluation of binding energy changes on mutation, (4) validation of protein stability, and (5) selection of rewiring mutations of interest.

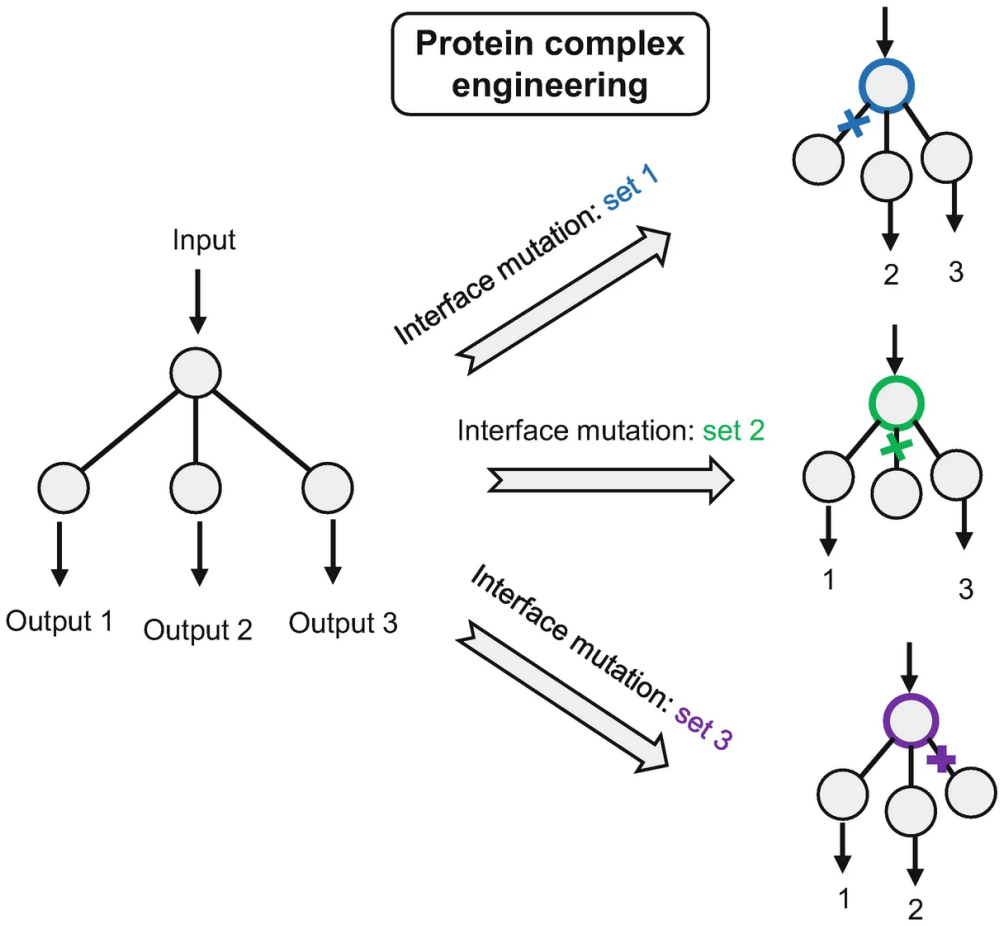

Figure 2.1 Illustration of rewiring mutations introduced in Ras binding to three different effectors. In each round, the goal is to disrupt binding of one of the effectors to Ras but keep binding of the other two effectors unchanged

2.2.3 Materials

Software

In this protocol, FoldX is used to calculate energies based on protein structures [74, 75]. It can calculate both Gibbs free energy of a protein (ΔG), evaluating its stability (*see* **Note 1**), and Gibbs free energy of binding (ΔG_{bind}), evaluating the interaction of two proteins in a complex (*see* **Note 2**). FoldX has two main functionalities; the first one is the evaluation of the energy of a protein using the FoldX force field. A force field is a set of theoretical and/or empirical energy terms based on which a total energy for a protein structure can be calculated (*see* **Note 1**). The second functionality is the adaptation of protein structures by exploring different structural configurations of the amino acid side chains, called rotamers. This can improve the energy of an experimentally determined structure. The inbuild rotamer library is also used for the actual step of in silico mutagenesis: FoldX can mutate all 20 natural amino acids. It has long been a difficult task to calculate absolute energies, and this task is still not adequately accomplished so far. Instead, relative energies are usually considered. This means that for every mutant structure of interest, a reference structure must be created, calculated, and compared to. This provides changes in Gibbs free energies ($\Delta\Delta G = \Delta G_{Mut} - \Delta G_{WT}$) and changes in Gibbs free energies of binding ($\Delta\Delta G_{bind} = \Delta G_{bind-Mut} - \Delta G_{bind-WT}$) (*see* **Note 3**). The FoldX software can be obtained on the official web page of the project: `http://foldxsuite.crg.eu/`. To install, please follow the instructions of the official manual.

Databases

FoldX requires protein structures as input. There are multiple ways of experimentally determining the structure of a protein (X-ray crystallography, NMR, cryo-EM), and experimentally determined protein structures have been invaluable in explaining the functions and mechanisms of action of proteins (*see* **Note 4**). All protein structures published are deposited in a common database, the Protein Data Base (PDB). There, the parameters of the experimental procedure

Table 2.1 Overview over the structures used in this example. Their PDB IDs as well as which structure is stored under which chain identifier are listed in this table

PDB ID	Chain A	Chain B	Reference
4K81	Grb14	Ras	[79]
3KUD	Ras	Raf (A85K)	[80]
6AMB	Ras	AF6	[81]

are available for download alongside the structures itself, all under a unique identifier, the PDB ID. The PDB can be accessed at `http://rcsb.org/`. For this example protocol, three protein structures from the PDB will be used (Table 2.1).

Scripts

All code examples given here are bash code. While the FoldX commands will be transferable without adaptation to the Windows command line or the MacOS console (see Note 5), the bash commands used (such as `grep`, `tr` or `mkdir`) are bash specific and their respective counterparts should be used. The full bash scripts describing the data acquisition for the three structures are available in the supplementary materials (Supplementary Files 1, 2, and 3). All data analysis and visualization were performed with R, ggplot2, and tidyverse [69, 82, 83]. A data analysis script for this protocol is available in the supplementary materials (Supplementary File 4).

2.2.4 Methods

Structure Preparation

Before working with the protein structures, there are some minor modifications to be made.

1 Download the complex structures of interest from the PDB. The PDB identifiers are 3KUD, 4K81, and 6AMB [79, 80, 81] (Table 2.1). The structures will

be downloaded in PDB file format. Place the structures in the current working directory and name them `[PDB ID].pdb`, i.e., `4K81.pdb`.

2 Prepare the structure for working with FoldX. Delete all crystal water and all unnecessary protein chains from the PDB file (*see* **Note 6**). Replace the character ‘ with ∗ for the correct parametrization of the GTP in the structure. This can be done manually, with a structure editor, or with text processing tool. As an example, this is the command for the structure 4K81, using simple bash tools:

```
# deletion of all HOH entries from the file
grep -v 'HETATM.*HOH' 4K81.pdb | \
# deletion of all ATOM, HETATM and TER entries for the chains C-H
    grep -v 'ATOM..................[C,D,E,F,G,H]' | \
    grep -v 'HETATM................[C,D,E,F,G,H]' | \
    grep -v 'TER...................[C,D,E,F,G,H]' | \
# replacement of ' with *
    tr "'" "*" > 4K81_clean.pdb
```

3 Repair of the structure with FoldX. During the repair of the structure, the amino acid side chains of the structure will be reoriented in order to minimize the energy of the structure. Energy calculations made with FoldX are more accurate after repairing a structure. The repair will create a report and a repaired structure `4K81_clean_Repair.pdb`. The command is:

```
foldx --command=RepairPDB --pdb=4K81_clean.pdb
```

4 Next, Ras-only structures will be obtained from the complex structures. This will be necessary for assessing the influence of Ras mutations on protein stability (Sect. 2.2.4). To do this, the effector chain will be deleted from the PDB file. For example, for the 4K81 structure:

```
# deletion of all ATOM, HETATM and TER entries for the chain A
grep -v 'ATOM..................[A]' 4K81_clean.pdb | \
grep -v 'HETATM................[A]' | \
grep -v 'TER...................[A]' > ./4K81_clean_Ras.pdb
```

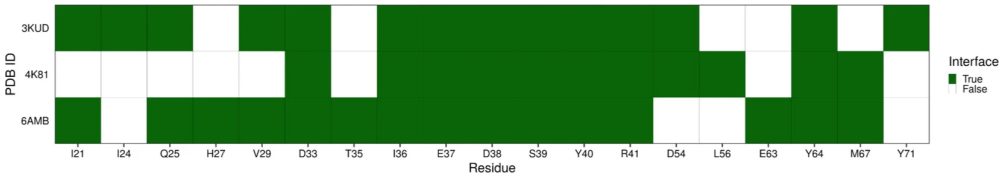

Figure 2.2 Interface residues determined by FoldX in all complex structures

5 Repair of the Ras-only structures with FoldX analogously to the repair of the complex structures.

Identification of Interface Residues

In order to reduce the computational cost, the mutations will be introduced only in residues of Ras that are participating in the interaction between Ras and its effectors, the so-called interface residues.

1 Analysis of the complex structure with FoldX. The command is:

```
foldx --command = AnalyseComplex --pdb = 4K81_clean_Repair.pdb
```

2 Extraction of the interface residues from the output file `Interface_Residues_4K81_clean_Repair_AC.fxout`. The residues are named in the following format: `[Amino Acid 1 Letter Code][Chain ID][Residue Number]`, i.e., `DB33` describes the aspartic acid at residue 33 of chain B.

3 Repeat for the other structures (3KUD, 6AMB) and compare the interface residues. For the following analysis, all interface residues from the three structures are pooled (*see* **Note 7**). Figure 2.2 shows the occurrence of interface residues in all three structures.

Evaluation of Binding Energy Changes on Mutation

In order to identify mutations in the interface of Ras and its effectors that selectively inhibit one interaction, a so-called position energy matrix for binding

energy $\Delta\Delta G_{bind}$ will be created for each of the complex structures. A position energy matrix is a matrix that contains all changes to the energy upon all mutations for all residues of interest.

1 Create an output directory, i.e., on a Linux operating system:

```
mkdir -p output_4K81
```

2 Run the PSSM command in FoldX to evaluate mutations in all amino acids of interest with regard to their influence on the binding energy between Ras and its effector. Amino acids of interest are all the interface residues identified for all three structures previously. FoldX expects amino acid inputs in the format `[Amino Acid 1 Letter Code][Chain ID][Residue Number][Mutation(s)]`, i.e., `IB21a` means that isoleucine at position 21 of chain B will be mutated to all canonical amino acids. Depending on the computational infrastructure this is run on, mutating 18 amino acids to all 20 canonical amino acids might take multiple hours. The command for 4K81 is:

```
foldx --command =Pssm \
    --pdb =4K81_clean_Repair.pdb \
    --output-dir = output_4K81 \
    --analyseComplexChains =A,B \
    --positions = IB21a,IB24a,QB25a,VB29a,DB33a,TB35a,IB36a, \
    EB37a,DB38a,SB39a,YB40a,RB41a,DB54a,LB56a,EB63a,YB64a,MB67a,
    \
    YB71a
```

While the PSSM command produces a lot of output files, the main information of interest can be found in `PSSM_4K81_clean_Repair.txt`. This file contains all changes of interaction energy of all mutations for all amino acids of interest in matrix format.

3 Repeat **steps 1** and **2** for 3KUD and 6AMB. A visualization of the results from all three structures can be found in Fig. 2.3.

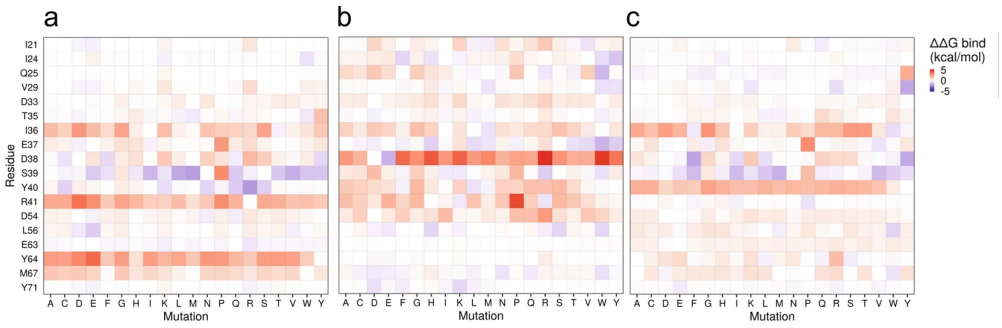

Figure 2.3 Heatmap of binding energy $\Delta\Delta G$ bind for all mutations for all residues. (a) Position energy matrix for 4K81. (b) Position energy matrix for 3KUD. (c) Position energy matrix for 6AMB

Validation of Ras Protein Stability

After determining the effect of mutations on binding energy, a necessary validation step is to check whether a mutation significantly influences the stability of our protein (Ras). To do this, we will create another position energy matrix investigating the changes of Gibbs free energy $\Delta\Delta G$ upon mutation of our residues of interest.

1 Create an output directory, i.e., on a Linux operating system:

```
mkdir -p output_4K81_Ras
```

2 Run the PositionScan command in FoldX. Similar to the PSSM command, it evaluates a list of mutations against the WT structure and generates a summarized output. In contrast to PSSM, which evaluates Gibbs free energy of binding (interaction), PositionScan evaluates Gibbs free energy (protein stability).

```
cd output_4K81_Ras
foldx --command=PositionScan \
    --pdb=4K81_clean_Ras_Repair.pdb \
    --pdb_dir=./.. \
    --positions=IB21a,IB24a,QB25a,VB29a,DB33a,TB35a,IB36a, \
    EB37a,DB38a,SB39a,YB40a,RB41a,DB54a,LB56a,EB63a,YB64a,MB67a,
   \
```

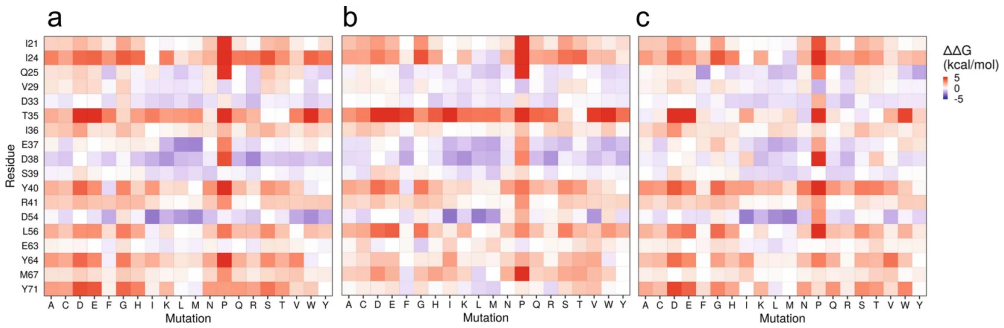

Figure 2.4 Heatmap of stability $\Delta\Delta G$ for all mutations for all residues. (a) Position energy matrix for 4K81. (b) Position energy matrix for 3KUD. (c) Position energy matrix for 6AMB

```
    YB71a
```

Again, many output files will be created. An overview over the Gibbs free energy changes can be found in `PS_4K81_clean_Ras_Repair_scanning_output.txt`.

3 Repeat **step 2** for 3KUD and 6AMB. A visualization of the results from all three structures can be found in Fig. 2.4.

Identification of Mutations of Interest

After obtaining both the changes to the binding energy $\Delta\Delta G_{bind}$ and the changes to the Gibbs free energy (stability) $\Delta\Delta G$ for all mutations of interest, the next step will be to select one or multiple mutations of interest. As stated earlier, the objective for this exercise is to generate mutants that specifically interrupt one Ras–Effector interaction, while simultaneously not affecting the other interactions. Therefore, the selection of mutations will be based on the following three criteria that have to be evaluated and balanced for each mutation:

1. Change of binding energy in the structure of interest $\Delta\Delta G_{bind-i}$: as high as possible.
2. Change of binding energy in the other structures $\Delta\Delta G_{bind-j}$: as close to zero as possible.

3. Change of Gibbs free energy of Ras structures $\Delta\Delta G$: as close to zero as possible.

While it would be possible to identify mutations of interest by simply studying the heatmaps, this is a complicated task. However, the selection process can be simplified and made objective by evaluating a scoring function based on the three criteria.

$$score = \Delta\Delta G_{bind-interest} - mean(|\Delta\Delta G_{bind-other}|) - mean(|\Delta\Delta G|)$$

The scoring function is implemented like this:

$$arg_{m \in M} max score(m) = a\left(\Delta\Delta G_{bind-i}(m)\right) - b\left(\frac{1}{n-1}\sum_{j \in X; j \neq i} |\Delta\Delta G_{bind-j}(m)|\right) - c\left(\frac{1}{n}\sum_{k \in X} |\Delta\Delta G_{k}(m)|\right)$$

with $\Delta\Delta G_{bind}$ being the changes to the binding energy upon mutation, $\Delta\Delta G$ being the changes to the Gibbs free energy upon mutation, $M = \{\texttt{I21A}, \texttt{I21C}, ..., \texttt{Y71Y}\}$ is the set of analyzed mutations, $X = \{\texttt{3KUD}, \texttt{4K81}, \texttt{6AMB}\}$ is the set of analyzed structures, n is the cardinality of the set X, and a, b, c being manually chosen weights for each of the three criteria.

Tables 2.2, 2.3, and 2.4 show the five top scoring mutations for each of the investigated structures according to our scoring function using the simple weights of $a, b, c = 1$. Considering the candidates in the light of our objective, there are multiple mutations that our approach identified as possible candidates. For the Ras-Grb14 structure (4K81), the mutations Y64L and Y64F, ranked first and third, respectively, seem promising candidates (Table 2). Both offer relatively high disruption of binding energy $\Delta\Delta G_{bind}$ (*see* **Note 8**), while simultaneously not impacting the stability $\Delta\Delta G$ (*see* **Note 9**).

The Ras-Raf structure (3KUD) is more challenging to evaluate. All mutations suggested affect the residue D38. This residue has been identified as a strong

Table 2.2 Overview over the top scoring mutations for the structure 4K81. The weights used in for the scoring procedure were $a = b = c = 1$. All energy values are given in kcal mol^{-1}

Mutation	Score	$\Delta\Delta G_{bind}$ 4K81	$\Delta\Delta G_{bind}$ 3KUD	$\Delta\Delta G_{bind}$ 6AMB	$\Delta\Delta G$ 3KUD	$\Delta\Delta G$ 4K81	$\Delta\Delta G$ 6AMB
Y64L	1.63	2.279	–0.013	–0.271	0.474	0.665	0.382
R41D	1.443	3.54	0.653	0.262	1.554	1.872	1.491
Y64F	1.228	1.432	–0.013	–0.106	–0.057	–0.332	–0.046
R41Q	1.145	1.949	1.259	0.042	0.185	0.185	0.089
Y64A	0.887	2.526	0.101	–0.047	0.506	2.586	1.601

Table 2.3 Overview over the top scoring mutations for the structure 3KUD. The weights used in for the scoring procedure were $a = b = c = 1$. All energy values are given in kcal mol^{-1}

Mutation	Score	$\Delta\Delta G_{bind}$ 3KUD	$\Delta\Delta G_{bind}$ 4K81	$\Delta\Delta G_{bind}$ 6AMB	$\Delta\Delta G$ 3KUD	$\Delta\Delta G$ 4K81	$\Delta\Delta G$ 6AMB
D38W	3.474	4.759	0.46	–0.235	–0.934	–1.083	–0.794
D38H	2.829	4.072	0.783	0.284	–0.526	–1.174	–0.429
D38R	2.114	5.607	1.816	0.998	–2.212	–2.5	–1.546
D38G	1.805	2.836	0.472	1.087	–0.041	–0.176	0.539
D38M	1.771	3.002	0.003	0.106	–1.58	–1.325	–0.622

Table 2.4 Overview over the top scoring mutations for the structure 6AMB. The weights used in for the scoring procedure were $a = b = c = 1$. All energy values are given in kcal mol^{-1}

Mutation	Score	$\Delta\Delta G_{bind}$ 6AMB	$\Delta\Delta G_{bind}$ 3KUD	$\Delta\Delta G_{bind}$ 4K81	$\Delta\Delta G$ 3KUD	$\Delta\Delta G$ 4K81	$\Delta\Delta G$ 6AMB
Y40M	1.112	1.685	–0.235	–0.037	0.359	–0.049	0.904
I36T	0.943	2.564	1.135	–0.319	0.54	1.128	1.015
Q25Y	0.821	2.143	0.163	0.089	–0.74	–1.024	–1.822
I36Q	0.553	1.835	0.95	1.229	0.115	0.396	–0.069
Y40F	0.544	1.34	–0.285	–0.053	–0.628	–0.793	0.461

contributor to the interaction of Ras and Raf, and in this particular structure, due to an interface mutation in Raf (A85K), the importance of D38 to this interaction has only been enhanced [80, 84]. This results in very high disruptions of binding energy $\Delta\Delta G_{bind}$ for mutations in D38, that dominate the scoring function (Table 2.3). While the mutations D38W and D38G, scoring first and fourth, respectively, are interesting candidates, another approach would be to adjust the weights in the scoring function such that $a < b = c$. This case highlights that the scoring function can and should be tailored to the problem at hand.

For 6AMB, the calculated scores are lower than for the other two structures (Table 2.4). Nonetheless, the mutations Y40M, Q25Y, and Y40F, scoring first, third, and fifth, respectively, seem promising.

2.2.5 Notes

1 The FoldX force field calculates the free energy of unfolding (ΔG, free energy difference between the folded and unfolded proteins) of a protein using the following equation [74, 75]:

$$\Delta G = a \times \Delta G_{vdw} + b \times \Delta G_{solvH} + c \times \Delta G_{solvP} + d \times \Delta G_{hbond} + e \times \Delta G_{wb} + f \times \Delta G_{el} + g \times \Delta G_{kon} + h \times \Delta S_{mc} + i \times \Delta S_{sc} + j \times \Delta G_{clash}$$

where ΔG_{vdw} is sum of Van der Waals contributions of all atoms, ΔG_{solvH} is the difference in solvation energy for apolar groups, ΔG_{solvP} is the difference in solvation energy for polar groups, ΔG_{hbond} is the free energy difference between the formation of an intra-molecular hydrogen-bond compared to inter-molecular hydrogen-bond formation with solvent, ΔG_{wb} is the extra stabilizing free energy provided by water molecules making more than one hydrogen bond to the protein, ΔG_{el} is the electrostatic contribution of charged groups, ΔG_{kon} is the effect of electrostatic interactions on the association rate constant (k_{on}), ΔS_{mc} is the entropy cost for fixing the backbone in the folded state, ΔS_{sc} is the entropic cost of fixing a side chain in a specific conformation, and ΔG_{clash}

is a measure of unfavorable steric overlaps between atoms in the structure. Parameters $(a, ..., j)$ are relative weights of the different energy terms.

2 For protein complexes, FoldX calculates the interaction energy (ΔG_{bind}) between two proteins, A and B, using the following equation [74, 75]:

$$\Delta G_{bind} = \Delta G_{AB} - (\Delta G_A + \Delta G_B)$$

3 Validation of energy changes calculated by FoldX with energies measured experimentally found that FoldX energy changes have a standard deviation of $\sigma = \pm 0.8\,\text{kcal mol}^{-1}$ [74]. When evaluating FoldX energies, it is important to keep in mind that these values are *in silico* calculations that come with a certain error.

4 FoldX works most reliable when high-resolution 3D structures are used, which are typically obtained from X-ray crystallography.

5 The FoldX software has been designed as a command line tool and has been released for all major operating system. As mentioned, the commands are transferable. Recently, a graphical interface to FoldX using the YASARA protein visualization program has been released [85, 86]. The protocol described here can also be replicated using the graphical interface; however, this protocol focusses on giving instructions for using the command line interface to FoldX. The FoldX plugin for YASARA along with instructions for the installation can be found on the official homepage of the FoldX project: `http://foldxsuite.crg.eu/`.

6 Only one biological assembly should be kept in the asymmetric unit, i.e., for the 4K81 structure, the chains C-H are removed. Sometimes, the biological assemblies are slightly different, so it might pay off to look through all of them.

7 Interface residues that do not contribute much to binding energy could nevertheless be good candidates for mutations as they might disrupt binding energy by introducing steric clashes in the interface.

8 A reasonable energy threshold for considering whether a mutation does disrupt a protein–protein interaction would be $\Delta\Delta G_{bind} > 1 - 2\,\text{kcal mol}^{-1}$. This would reduce the binding affinity of an interaction by a factor of 5–10, which would be enough to dramatically reduce the number of complexes formed. This is especially true in a highly competitive system such as the Ras–Effector system.

9 Disruption of stability should be evaluated in a different way from disruptions of binding energy. In general, proteins are more resistant to mutations affecting stability. A reasonable energy threshold would be a disruption of $\Delta\Delta G < 1.6\,\text{kcal mol}^{-1}$, at which proteins would be affected. This threshold of twice the standard error of FoldX is commonly applied (*see* **Note 3**).

2.2.6 Acknowledgment

This work is part of the research program “Quantitative and systems analysis of (patho)
physiological signalling networks” with project number 16/FRL/3886, which is financed by Science Foundation Ireland (SFI) (to C. Kiel).

2.3 Discussion

In this paper, we presented the theoretical background of the *in silico* design of network-rewiring mutations using FoldX. These mutations are also called "branch-pruning" mutations.

The design of network-rewiring mutations *in silico* has the benefit that it is possible to check all possible mutations with low cost and time investment. However, the *in silico* predictions should only be seen as a first step, and experimental validation of interesting candidates is recommended.

For this work, we were using FoldX for the energy calculation on the protein structures. FoldX is a well-established method for the assessment of point mutations on protein stability and protein-protein interactions [74, 75, 87]. However, other methods for determining the energetic impact of point mutations such as Rosetta [88] could be used analogously. For all of these methods, it is important to note that the quality of the protein structures used is a major factor on the expected accuracy of the results.

The work in this paper lays the foundation for the selection and evaluation of "branch-pruning" mutations analysed in chapter 4.

3 HOMELETTE: a unified interface to homology modelling software

3.1 Introduction

Protein structures are useful tools for understanding protein function on the molecular level and for drug design. Therefore, predicting the structures of proteins for which no structure has been experimentally identified so far is of major interest.

Homology modelling is a method for the computation prediction of protein structures. Briefly, the idea is that proteins that are closely related on the level of amino acid sequence will also share strong similarity on the structural level. Using information about the structures from close relatives in combination with sequence alignments allows then the construction of computational models.

Over the last decades, a significant number of different software for model generation and model evaluation has been developed by different groups [89, 93, 94, 95, 96, 97, 98, 99, 100, 90, 91, 92]. In the following paper, I am presenting a Python package called HOMELETTE that creates a unified interface to many of these software packages. This makes it easy to assemble and benchmark custom modelling pipelines.

3.2 HOMELETTE: a unified interface to homology modelling software

Philipp Junk[1,*], Christina Kiel[1]

[1] Systems Biology Ireland and UCD Charles Institute of Dermatology, School of Medicine, University College Dublin, Dublin 4, Ireland

[*] Corresponding author

The Manuscript was **published** in **Bioinformatics** on the 15/03/2022.
It is available at `https://doi.org/10.1093/bioinformatics/btab866` under an Open Access license.

The content is identical to the version published, with changes to formatting. The supplementary materials, which are currently not available on the publisher's web site due to technical issues, are included as well.

3.2.1 Abstract

Summary

Homology modelling, the technique of generating models of 3D protein structures based on experimental structures from related proteins, has become increasingly popular over the years. An abundance of different tools for model generation and model evaluation is available from various research groups. We present HOMELETTE, an interface which implements a unified programmatic access to these tools. This allows for the assemble of custom pipelines from pre- or self-implemented building blocks.

Availability and implementation

HOMELETTE is implemented in Python, compatible with version 3.6 and newer. It is distributed under the MIT license. Documentation and tutorials are available at Read the Docs (`https://homelette.readthedocs.io/`). The latest version of HOMELETTE is available on PyPI (`https://pypi.org/project/homelette/`) and GitHub (`https://github.com/PhilippJunk/homelette`). A full installation of the latest version of HOMELETTE with all dependencies is also available as a Docker container (`https://hub.docker.com/r/philippjunk/homelette_template`).

Supplementary information

Supplementary data are available at Bioinformatics online.

3.2.2 Introduction

Access to homology modelling tools has become increasingly simpler over the last years. There is a multitude of web services such as SWISS-MODEL offering total automation of the whole process. These are great tools for small homology modelling projects [101]. However, medium to large scale projects, aiming to model the structures of tens or hundreds of proteins with different homology modelling software in a full- or semi-automated manner are faced

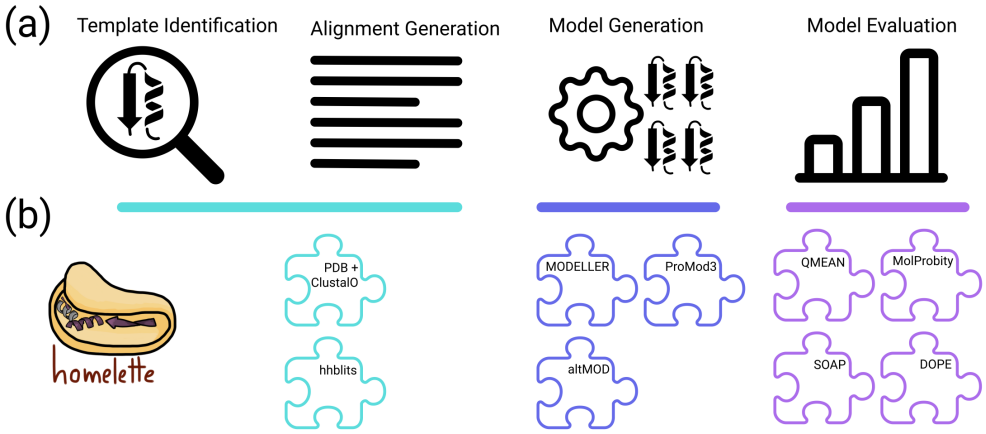

Figure 3.1 Homology modelling pipeline. (a) General pipeline of homology modelling from left to right. (b) Building blocks implemented in HOMELETTE and how they correspond to the steps in homology modelling

with a very tedious exercise. Most of the popular homology modelling services offer command line tools. However, these tools come with different interfaces and work with different file types. The same is true for software aiming to evaluate homology models.

The general flow of a homology modelling pipeline is depicted in Figure 3.1a [96]. Usual requirements for most homology modelling software are a multiple sequence alignment (MSA) of the target sequence against one or multiple template sequences, as well as template structures. Using the information from the alignment and the template structure(s), a homology modelling algorithm assembles one or multiple models. Afterwards, these are evaluated by some evaluation metrics in order to select the best model(s).

Exchanging components of the pipeline such as the modelling algorithm or the evaluation metrics is not trivial due to the problems outlined above. Therefore, the motivation behind HOMELETTE is to provide a modular homology modelling interface that can be used to construct pipelines with diverse modelling and evaluation tools within the same interface. The focus is also on making it easy for the user to implement new building blocks that fit into the

framework. This interface can be used to easily assemble custom pipelines and streamline medium to large scale homology modelling projects (Fig. 3.1b).

3.2.3 Implementation

The HOMELETTE interface is fully implemented in Python. Python is a popular and accessible programming language extensively used in the scientific community [102].

HOMELETTE is built with modular design principles in mind. Template identification/alignment generation, model generation and model evaluation are designed as interchangeable building blocks that interact with the other components of the pipeline in an identical manner. This allows for the easy assembly of custom pipelines by freely combining these building blocks. Alignment generation and template processing building blocks are available for identifying templates with the RCSB Search Web API using MMseq2 [103, 104] and align them with Clustal Omega [105, 106], or using HHSuite3 [107]. Model generation building blocks are currently available for MODELLER [89, 96], altMOD [97] and ProMod3 [98, 99]. Model evaluation building blocks are available for DOPE scores [100], SOAP scores [90], QMEAN [91, 92], QMEAN DisCo [93] and MolProbity [94, 95]. A good model is expected to have a low DOPE score, a low SOAP score, a high QMEAN score and a MolProbity score as close to 0 as possible. A list of the implemented building blocks is available in Supplementary Table 3.S1.

In addition, new building blocks can be implemented and seamlessly fit into existing pipelines allowing for even further customization. This is particularly useful for integrating software for which no building block is available yet into the framework. Users are strongly encouraged to share their custom building blocks with the community, and an extension framework has been set up to make this possible.

Extensive documentation and tutorials teach the user how to use these building blocks, how to implement new building blocks and how to assemble them into more complex pipelines. The documentation is available online at `https://homelette.readthedocs.io/`. The tutorials are hosted together with

the documentation, or as interactive Jupyter notebooks on the GitHub page and in the Docker container.

HOMELETTE does not have any model building or model evaluating capacities on its own, but its strength comes from the integration of different software. Due to these design choices, it is reliant on third-party software (Supplementary Table 3.S1). All currently integrated software is freely available for academic research. The documentation gives instruction on how to acquire and install third-party software. Alternatively, HOMELETTE is also available as a Docker container with all third-party software already installed.

3.2.4 Application

As an example for the custom assembly of alignment generating, homology modelling and model evaluation building blocks into custom pipelines, the ARAF protein was modelled (Supplementary Fig. 3.S1). Starting from the sequence, the templates 3NY5 (BRAF) and 4G0N (RAF1) were identified, aligned and processed. In order to show how different modelling building blocks can be used interchangeably, two MODELLER building blocks with different parameters for model refinement were used. Evaluation was performed by using SOAP scores and MolProbity scores, which were summarized to a combined score using Borda count (Supplementary Fig. 3.S1b). As expected, the modelling routine that spends more time on model refinement generates better models. There are also differences between the templates to be observed. The code to execute this example as well as to generate the visualization is made fully available in Tutorial 7.

3.2.5 Conclusion

There are three major determinants for the quality of a homology model. These are the alignment used, the quality of the template structures and the algorithm chosen for generating the models [96]. HOMELETTE leaves the selection of all three determinants in the hand of the user. The user has agency which

modelling software to use and compare, as well as full control over generating and refining the alignment and selecting templates.

We explain and demonstrate the use of HOMELETTE in the series of eight tutorials. The tutorials culminate in a tutorial about pipeline assembly, which has been shown as an example pipeline for a proof of concept in this publication (Supplementary Fig. 3.S1).

In conclusion, HOMELETTE offers a unified, simple and well-documented interface to a multitude of popular homology model and model evaluation software. Its modular design principles allow users to assemble their own pipelines in an easy and consistent manner. Simple implementation and extensive documentation make it possible to extend HOMELETTE with other software, while retaining the same programmatic interface. This gives users even more freedom to assemble the best custom pipeline for their particular project. This could prove useful for large scale projects such as the structural modelling of whole biological systems.

3.2.6 Acknowledgements

The authors acknowledge all scientists that distribute their work free of charge or Open Source for making this project possible. They also thank all members of the Kiel lab for discussions and critical reading of the manuscript.

Funding

This work is part of the research program “Quantitative and systems analysis of (patho)
physiological signaling networks” [16/FRL/3886], which is financed by Science Foundation Ireland (SFI) to C.K.

Conflict of Interest

None declared.

Data availability

No new data were generated or analysed in support of this research.

3.2.7 Supplement

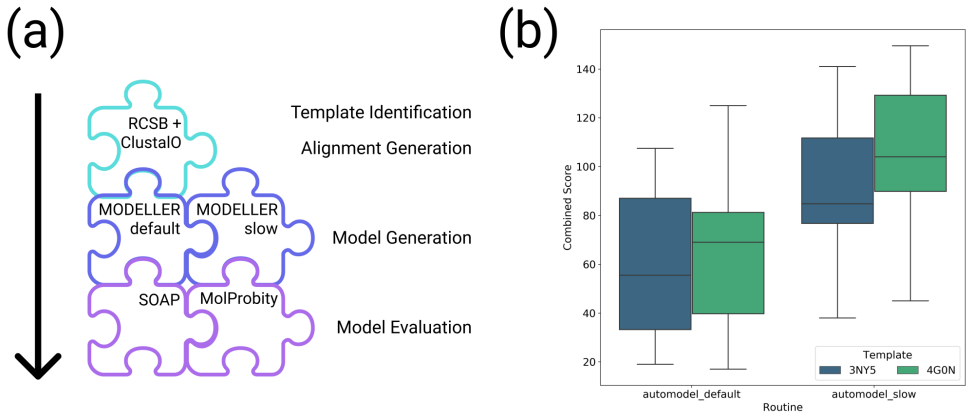

Figure 3.S1 (a) Example of a modelling pipeline. For the query sequence (ARAF RBD in this case), templates were identified using the RCSB Search API, and an alignment was generated using Clustal Omega. 3NY5 and 4G0N were selected as templates. Models were generated using the MODELLER homology modelling software with two different parameter sets. The models were evaluated using SOAP scores and MolProbity scores. (b) The combined scores of the models generated by the pipeline described in (a). Scores were combined by using Borda count. A higher score indicates a better model. The full code of this example is available as Tutorial 7 (`https://homelette.readthedocs.io/en/latest/Tutorial7_AssemblingPipelines.html`).

Table 3.S1 Building Blocks currently implemented in HOMELETTE

Building Block	Type	Program/Software	Reference
AlignmentGenerator_pdb	Alignment generation, Template processing	RCSB Search API (MMseq2), Clustal Omega	[103, 104, 105, 106]
AlignmentGenerator_hhblits	Alignment generation, Template processing	HHSuite3	[107]
AlignmentGenerator_from_aln	Alignment generation, Template processing	-	-
Routine_automodel_default	Model generation	MODELLER	[89, 96]
Routine_automodel_slow	Model generation	MODELLER	[89, 96]
Routine_altmod_default	Model generation	altMOD	[97]
Routine_altmod_slow	Model generation	altMOD	[97]
Routine_promod3	Model generation	ProMod3	[98, 99]
Routine_complex_automodel_default	Model generation	MODELLER	[89, 96]
Routine_complex_automodel_slow	Model generation	MODELLER	[89, 96]
Routine_complex_altmod_default	Model generation	altMOD	[97]
Routine_complex_altmod_slow	Model generation	altMOD	[97]
Routine_loopmodel_default	Model generation	MODELLER	[89, 96]
Routine_loopmodel_slow	Model generation	MODELLER	[89, 96]
Evaluation_dope	Model evaluation	MODELLER	[100]
Evaluation_soap_protein	Model evaluation	MODELLER	[90]
Evaluation_soap_pp	Model evaluation	MODELLER	[90]
Evaluation_qmean4	Model evaluation	QMEAN	[91, 92]
Evaluation_qmean6	Model evaluation	QMEAN	[91, 92]
Evaluation_qmeandisco	Model evaluation	QMEAN	[91, 92, 93]
Evaluation_mol_probity	Model evaluation	MolProbity	[94, 95]

3.3 Discussion

With the release of AlphaFold2 in 2021, the landscape of protein structure prediction has been changed [9, 59]. In particular in cases when no closely related structures are available, AlphaFold2 vastly outperforms homology modelling methods. Nonetheless, homology modelling is a useful technique for the *in silico* generation of models. In my experience (compare also with 4), homology modelling approaches still perform better than AlphaFold2 for the modelling of related structures around a well-defined interface. Additionally, some homology modelling software is able to work with heteroatoms such as small molecules, co-factors or ions, which is currently not supported. So while homology modelling methods will be superseded by deep learning based methods (AlphaFold2 and related methods) in the long run, currently there is still a use for homology modelling.

The HOMELETTE package presented in this paper is an interface to other popular and well-established homology modelling and model evaluation software, such as Modeller and ProMod3. In particular, it is useful for the creation and comparison of different pipelines, which can be assembled from building blocks. Other software can easily be implemented as custom building blocks and integrated in existing pipelines. HOMELETTE is well documented and offers an associated Docker container for easy set-up.

4 Structure-based prediction of Ras-effector binding affinities and design of “branchegetic” interface mutations

4.1 Introduction

As outlined in the introduction, one of the main aims of the work done in my PhD was the investigation of the interaction of RAS with its effector proteins. While some of these interactions are well-characterized on the biological or the structural level, there are a number of proteins in the human genome that encode domains that would potentially allow them to interact with RAS. The aim of the following work is to explore all potential RAS effector interactions systematically.

For this, a structural biology approach was used. The paradigm of analysing protein structure to inform biological systems is not new. However, with the recent release of AlphaFold2 addressing the protein structure prediction problem, and subsequent technological development, new options are available in the structural biology space.

The RAS effector system is a well-suited system to test structural biology approaches. There is a massive amount of data available for this system, which can be used to inform, train, and evaluate the modelling approaches. For example, there are multiple structures of RAS in complex with different effectors available in the PDB, making it possible to investigate the binding

mode of these domains. Similarity, binding affinities for multiple effectors to RAS have been measured experimentally.

In the following paper, computational structural biology approaches were used to generate structural models of RAS in complex with effectors, as well as mathematical models of the competition of effectors for the RAS interface in different tissues. Potential interface mutations were selected from the structural models and their impact on the systems model were evaluated.

4.2 Structure-based prediction of Ras-effector binding affinities and design of “branchegetic” interface mutations

Philipp Junk[2,3,*], Christina Kiel[1,2,3]

[1] Department of Molecular Medicine, University of Pavia, 27100 Pavia, Italy

[2] Systems Biology Ireland, School of Medicine, University College Dublin, Dublin 4, Ireland

[3] UCD Charles Institute of Dermatology, School of Medicine, University College Dublin, Dublin 4, Ireland

[*] Corresponding author

This manuscript in currently **in press** at **Structure**. A preprint is available at `https://doi.org/10.1101/2022.09.04.506480`.

The content is identical to the version available on BioRXiv, with changes to formatting and minor changes to bibliography.

4.2.1 Summary

Ras is a central cellular hub protein controlling multiple cell fates. How Ras interacts with a variety of potential effector proteins is relatively unexplored, with only some key effectors characterized in great detail. Here, we have used homology modelling based on X-ray and AlphaFold templates to build structural models for 54 Ras-effectors complexes. These models were used to estimate binding affinities using a supervised learning regressor. Furthermore, we systematically introduced Ras "branch-pruning" (or branchegetic) mutations to identify 200 interface mutations that affect the binding energy with at least one of the model structures. The impacts of these branchegetic mutants were integrated into a mathematical model to assess the potential for rewiring interactions at the Ras hub on a systems level. These findings have provided a quantitative understanding of Ras-effector interfaces and their impact on systems properties of a key cellular hub.

4.2.2 Introduction

Ras is a key cellular signaling hub and oncogene [108]. The first correct Ras structure was famously described in 1989 [37, 109] and consists of the G domain super-fold (six β-strands and five α-helices; [108]). It's main structural and functional characteristic is the nucleotide binding site (the typical α,β-fold of nucleotide binding proteins), which can bind GTP and hydrolyze it to GDP. Then, GDP gets released and a new nucleotide (favorably GTP since it is higher abundant in cells) gets bound. Depending on which nucleotide is bound, the functional state of RAS is different. GDP bound Ras assumes a so-called inactive conformation, whereas upon the binding of GTP, the conformation of Ras is called active. The difference between the active and the inactive conformation is that, in the active conformation, two loop regions called switch 1 and switch 2 are interacting with the GTP and thereby tightly bound to it [108] (Figure 4.1A). This rearranges the interface of Ras in a specific way so that effector proteins with a specific structural motive can interact with Ras. The structural motive required for the interaction with Ras has a ubiquitin

domain super-fold [40]. There exist three families of these domains, called Ras binding domain (RBD), Ras association domain (RA domain) and PI3K-Ras binding domain (PI3K RBD) with only minor sequence and structural differences between them. In the following, for the sake of simplicity, all of these will be called RBDs.

Ras-GTP can interact and activate many downstream effectors, some of which are better studied than others [33]. Well-studied effectors include PI3-kinases, RalGDS, and Raf kinases, which control important cellular processes such as survival, polarization, adhesion, migration, and proliferation. In our previous work, we characterized a set of 56 RBD-containing proteins as potential Ras effectors, which converge into 12 classes that are linked to different downstream cellular processes and phenotypes [30, 15, 29, 57].

Ras-effector interactions are interesting from a systems biology and network point of view. Effectors use a mutually exclusive binding site on Ras, hence competition for binding can occur under certain conditions [110]. Also, the experimental and predicted affinities between RBDs and Ras-GTP vary, ranging from nano- to micromolar K_d values [29]. Previously we analyzed how the amount of effectors in complex with Ras varies in different cell types [30, 15]. We find that only 9 effectors are predicted to bind in significant amount to Ras-GTP using the RBD alone [15]. However, 31 effectors are predicted to form significant complexes with Ras if they are additionally recruited to the plasma membrane via other domains present in effectors (piggyback mechanism, [111]). Hence, even weak binding affinities on the level of RBD-RAS binding can turn into significant complex formation if effectors are recruited (e.g. in a context specific manner). Thus, for systems analyses and computational models, we need good estimations for all binding affinities between RBDs and Ras.

Structural analysis can lead to a deeper understanding of protein-protein interaction (PPI). Previous we used homology modelling, based on experimentally determined complex structures, to predict binding between effectors and Ras-GTP [49, 50]. While this has provided insights into binding propensities, a limitation of this work was that no quantitative binding affinities were predicted, but qualitative binding by classifying effectors into categories of

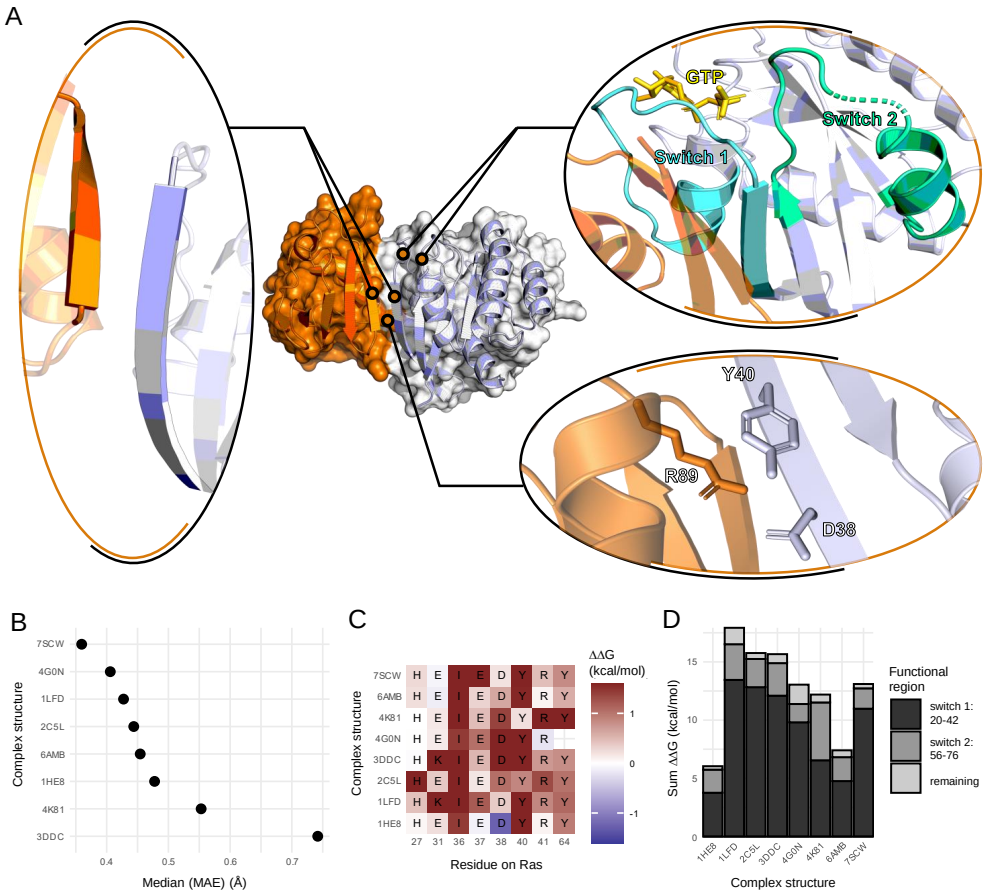

Figure 4.1 Noticeable features of the Ras-RBD interfaces (A) Overview of interface features for the Ras-RAF1 interface. Highlighted are the intramolecular β-sheet alignment, the assembly of the Ras interface by the switch regions and energetic hotspots in the interface. (B) β-sheet alignment for experimentally determined complexes (MAE). A lower value indicates a more similar alignment of the intramolecular β-sheets compared to other crystal structures. (C) FoldX alanine scan hotspots on RAS for experimentally determined complex structures. The color scale is confined to the limits [–1.6 kcal mol^{-1}, 1.6 kcal mol^{-1}]. Positive ΔΔG values indicate a loss in binding energy upon mutation, which reflects that the respective amino acid contributes to binding. (D) Energetic contributions of functional regions based on alanine scan analysis.

"binding", "non-binding", and "twilight". It is now timely to revisit homology modelling of Ras-RBD complexes, as new template structures are available by (i) X-ray crystallography (8 of 56 effectors are crystallized in complex with RAS) and (ii) AlphaFold [9, 59]).

Ras is frequently mutated in different cancers, where aberrant Ras signaling plays a role in cancer initiation, progression, and metastasis via alterations in metabolism, proliferation, and survival [112]. As Ras cannot be directly targeted (or only certain Ras mutations) [113], much hope rests on network-centric approaches [114] that involve Ras-effector interactions or downstream pathways [57]. Therefore, tinkering with Ras-effector binding is an attractive alternative strategy for finding suitable targets for therapeutic interventions. Previously we developed a "branch-pruning" (or "branchegetic", in analogy to "edgetics" [22] and "enedgetics" [23]) strategy for Ras effector interactions, where mutations are introduced into Ras that differentially impact binding to effectors [115]. For example, introducing a mutation can result in a steric clash in the interface formed with one effector, but not with another; hence the interaction with one effector is broken while intact for another.

In this work, we first generate homology models of all Ras-effector (RBD) interactions and predict affinities of the RBDs in complex with Ras-GTP. We then use the generated model structures to employ a systematic branchegetic strategy that explores the impact of Ras interface mutants on binding to all effectors. Altogether, our results contribute to a quantitative and systems-level understanding of Ras-effector interactions and further our understanding of Ras in health and disease.

4.2.3 Results

Structural analysis of experimental complex structures

From the few experimentally determined complex structures of Ras with these RBDs, there are some structural similarities that can be determined. The main interface on the site of Ras consists of the two switch regions, switch 1 and switch 2. One of the main structural features of the interface between Ras

and RBDs is the formation of an intra-molecular β-sheet between β2 on RAS and β2 on the RBD. This interaction is a highly conserved structural feature across all available complex structures, with deviations in orientation of less than 1 Å (Figure 4.1B, see Methods for MAE). Analyzing the energetic profile of the interface *in silico* using the FoldX force field shows that there are also some recurring hotspots in the interface (highlighted in Figure 4.1C). I36, D38 and Y40 are well characterized as important residues for the interaction between Ras and effector domains. Additionally, for the structures 3DDC and 1LFD, the mutation E31K was introduced to stabilize the complex for crystallization. Our analysis confirms that this interface mutation has indeed been favorable for the complex formation. Finally, the energetic contributions of the function regions switch 1 and 2 to the interface were analyzed. Both relative and absolute contributions are diverse, although switch 1 contributions to binding dominate (Figure 4.1D). Altogether, the analysis of existing Ras-RBD effector structures indicates that although there are many common features of the Ras RBD interface such as the intra-molecular β-sheet or the hydrophobic patch around I36, the actual energetic contributions can come from different parts of the interface. It also sets the basis for a successful homology modelling approach.

Homology modelling and characterization of modelled interfaces

In order to study these interface features in a more diverse set of structures, homology models of the complex between RAS and RBD were constructed for all proteins containing an RBD in the human proteome.

The homology modelling pipeline is based on the already existing complex models (Table 4.S1). Also, with the recent release of AlphaFold2 [9] and the accompanying AlphaFold Protein Structure Database [59], the RBD domains of all potential effectors were extracted from that database and used. The structures of RBDs are predicted with good confidence by AlphaFold2 and our analysis indicates that AlphaFold2 is reliable at predicting the RBD fold (Figure 4.S1). Additionally, AlphaFold2 complex modelling was attempted for all potential complex structures and models which AlphaFold2 were confident in (by Predicted Alignment Error (PAE)) and where the β-sheet alignment of the

interface was within a tolerance to what has been observed in crystal structures, were used as templates as well (Figure 4.S2 and 4.S3, compare MAE Figure 4.1B). There are two kinds of templates to use: 1) complex templates which comprised of the already experimentally determined complex structures of Ras and RBDs, as well as "good" AlphaFold2 predicted complex templates (Table 4.S1), and 2) templates of the RBDs alone which were extracted from the AlphaFold2 EMBL database (Table 4.S2). An overview over the pipeline is depicted in Figure 4.2. Homology modelling was performed using homelette [116] with modeller and altMod. Evaluation of predicted structures was performed using QMEAN, MolProbity and SOAP potentials. The top 300 models for each target for each source of complex templates (experimental or AlphaFold2) were selected by combined score and FoldX analysis (interaction energy and alanine scan) was performed.

In order to evaluate our approach, we generated a validation set, in which we created models for the structures already solved by X-ray crystallography without using information from the specific structure. Models generated in this validation set were compared to the underlying ground truth by assessing their correlation of the *in silico* alanine scan results to those of the crystal structures of interest. Using this ground truth, different methods to select representative models from the hundreds of structures were assessed. In particular, we evaluated the hyperparameters of an unsupervised learning pipeline comprised of different feature selection and dimensionality reduction strategies followed by clustering with the OPTICS algorithm (see Methods for more details about the hyperparameter space). After clustering, three representative structures were chosen. Based on the performance in the validation set, the optimal set of hyperparameters was chosen (Table 4.S3, see Figures 4.S4 and 4.S5). The described strategy for identifying representative structures was then applied to all target complexes and we ended up with three representative complex structures for each effector.

Analogous to how we characterized the interfaces of the experimentally determined complex structures before, we performed the same analysis on the complex models. The overall FoldX interaction energies for the models are diverse, indicating that maybe some of the complexes are energetically

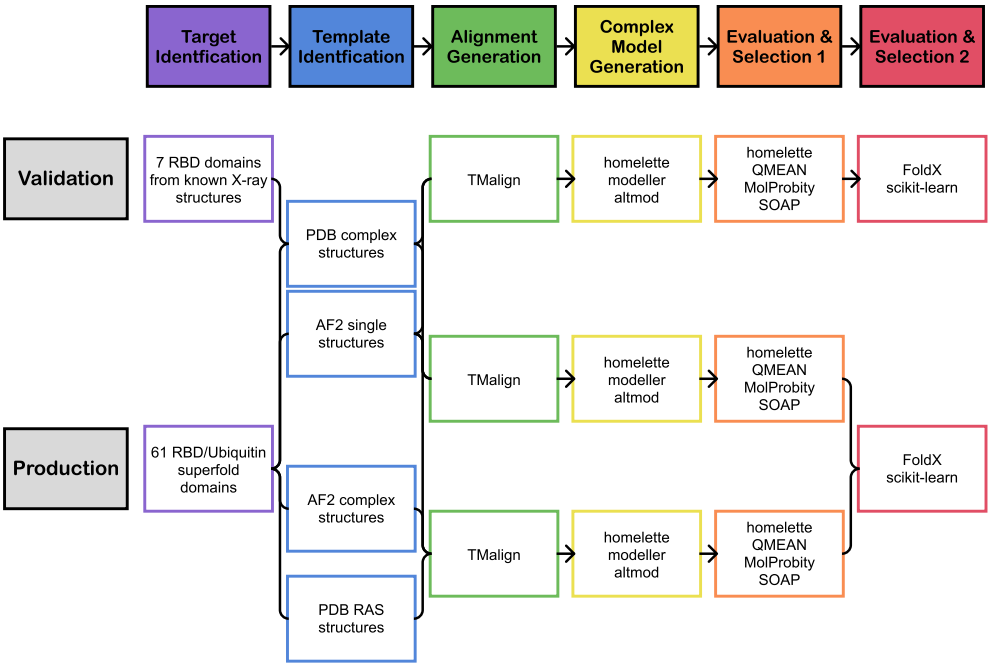

Figure 4.2 Overview over the homology modelling pipeline The principal homology modelling steps are shown from left to right with the steps of “Target Identification”, “Template Identification”, “Alignment Generation”, “Complex Model Generation”, “Evaluation & Selection 1”, and “Evaluation & Selection 2”. In the “Validation” process, the homology modelling pipeline is run on a set of 7 RBD domains from known x-ray structures. The “Production” process describes the generation of 61 domains (RBD and Ubiquitin super-fold domains as controls).

unfavorable and would not form (Figure 4.S6B). In general, the binding energies are lower than what would be observed in crystal structures, which is to be expected. There are one or two outliers with regards to FoldX binding energy, namely RASSF8 and PIK3C2B. In particular, RASSF8 is also showing an uncommon hotspot profile, with multiple unfavorable hotspots that are only appearing for this set of structures (Figure 4.S6A). Based on this behavior, RASSF8 is excluded from further analysis.

The analysis of hotspots confirmed the already established hotspots. I36, Y40, D38 and E37 are the most commonly observed hotspots (Figure 4.3A, Figure 4.S6A). Interestingly, while I36, Y40, and E37 are exclusively favorable to the interaction with the effector protein, D38 seems to be also unfavorable in some of the structures (Figure 4.3B). The energetic diversity of the hotspot D38 was further investigated in the models. For this, two models were picked for which D38 was a favorable hotspot in the alanine scan analysis (Figure 4.3C: RASSF1, Figure 4.3D: RGL3), and two models were picked for which it was unfavorable (Figure 4.3E: ARAP2, Figure 4.3F: RAPGEF3). Next, neighboring amino acids were analyzed. For favorable interactions, we were able to observe positively charged amino acids. On RASSF1, we find K216 and H217, whereas on RGL3, there is R630. These amino acids probably form strong interactions with the negatively charged D38 on Ras. In contrast, for the models where D38 comes up as an unfavorable hotspot in the interface, we observe an uncharged, mostly hydrophobic neighborhood.

Estimation of binding energies

One of the applications of the structural models that were generated was to use them for the estimation of binding energies. Since experiments measuring the binding energy between two proteins are experimentally very challenging and error-prone [117], we were implementing an *in silico* approach. Also, while FoldX is good at predicting energy changes to interaction energy or protein stability on mutation, the absolute interaction energies for protein complexes usually are not well correlated with experimental values [75]. Because of this, a supervised learning regression pipeline was built based on features extracted

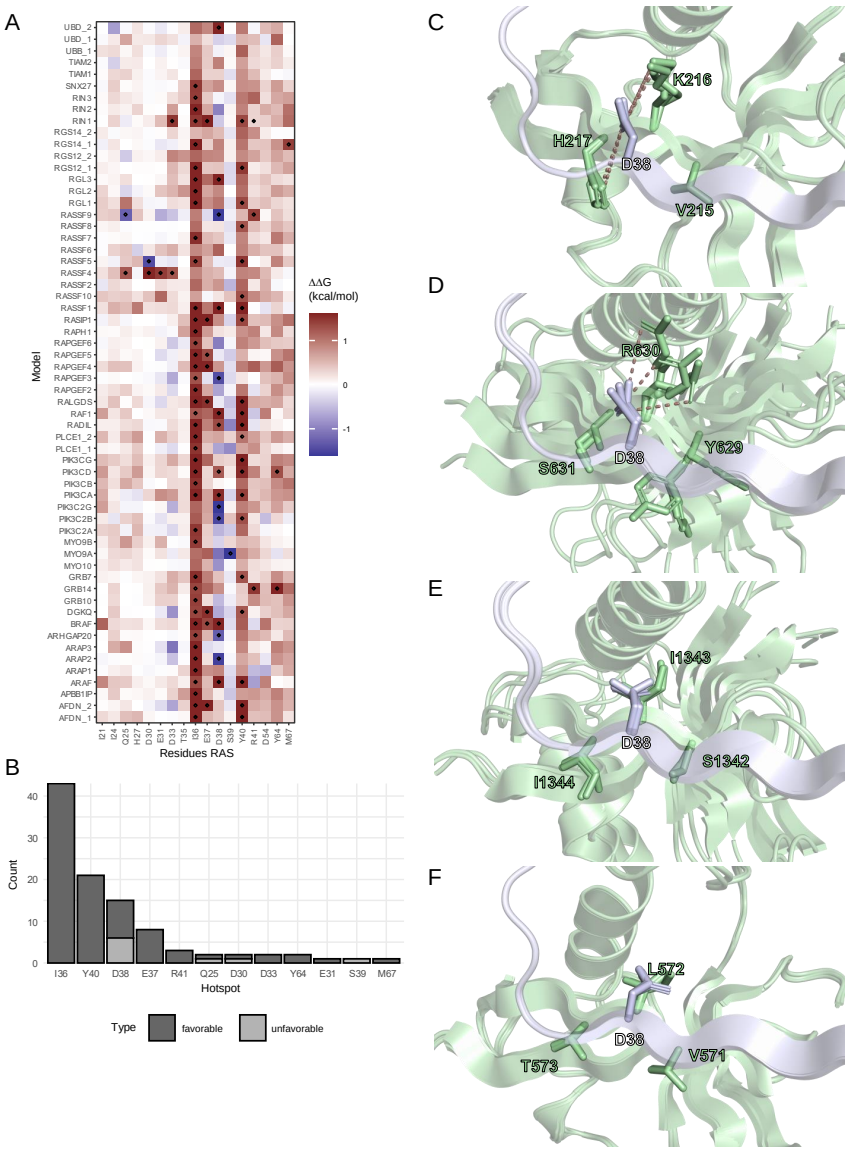

Figure 4.3 Energy hotspots in modelled structures (A) Heatmap of FoldX alanine scan averaged from the three representative structures on each target. The color scale is confined to the limits [–1.6 kcal mol^{-1}, 1.6 kcal mol^{-1}]. Hotspot residues with a $\Delta\Delta G \geq 1.2$ kcal mol^{-1} or ≤ -1.2 kcal mol^{-1} were marked. C-F) Local neighborhood of D38 in structures where D38 is a favorable hotspot (C: RASSF1, D: RGL3) or an unfavorable hotspot (E: ARAP2, F: RAPGEF3). Ras and the effector structures are visualized in blue and green, respectively. Polar interactions between charged amino acids are indicated with dashes.

from the modelled structures and a collection of experimentally determined binding energies of different Ras-effector complexes (Table 4.S5). From a combination of different regressors, feature extraction procedures, and hyperparameters, the best approach was determined using a cross validation strategy (see methods). A support vector machine-based regressor (see hyperparameters in Table 4.S4) performed best in cross-validation with an R2 score of 0.53. Then, the performance of the best approach was evaluated in an out-of-sample test set, where it achieved an R2 score of 0.77. The model was then used to predict the interaction energies for the complexes without prior experimental measurements (Figure 4.4, Table 4.S5). The predicted binding affinities for our models range 4 orders of magnitude, between the highest predicted affinity for BRAF of 0.02 µM to PIK3C2B with the lowest predicted affinity of 588 µM. The highest binding effectors are quite well characterized (RAF family, PI3K, RASSF5, RIN1, RalGDS, AFDN [29]). As experimental measurements of the PI3K family members are difficult as the RBD is not easy to express and purify in isolation, it is noteworthy that we assign three of the good binding affinities to PI3K family members (PIK3CA, PIK3CD, PIK3CG). A big group of effectors has affinities in the range of 1 to 10 µM. For example, RASIP1 was previously in the "likely no binding" category [29], and is now predicted to have an affinity in complex with Ras of 2.7 µM. RAPGEF5 and RGL3 were previously in the "unknown" [29] category and have predicted affinities of 9.2 and 5.6 µM, respectively. Another big group of effectors has affinities in the range of 10 to 100 µM. Especially the first one could be interesting for modulation of binding affinity by the piggy back mechanism [15].

Switch contributions to binding

Having Ras-effector structural models available allowed us to analyze the individual contributions of switch 1 and switch 2 to binding using the results from alanine scanning (similarly as done before for the X-ray Ras-effector structures). We find that generally most structures are dominated by favorable switch 1 contributions (Figure 4.5 and Figure 4.S7). Switch 2 contributions are surprisingly small. We also predict more contributions from regions outside switch 1 and

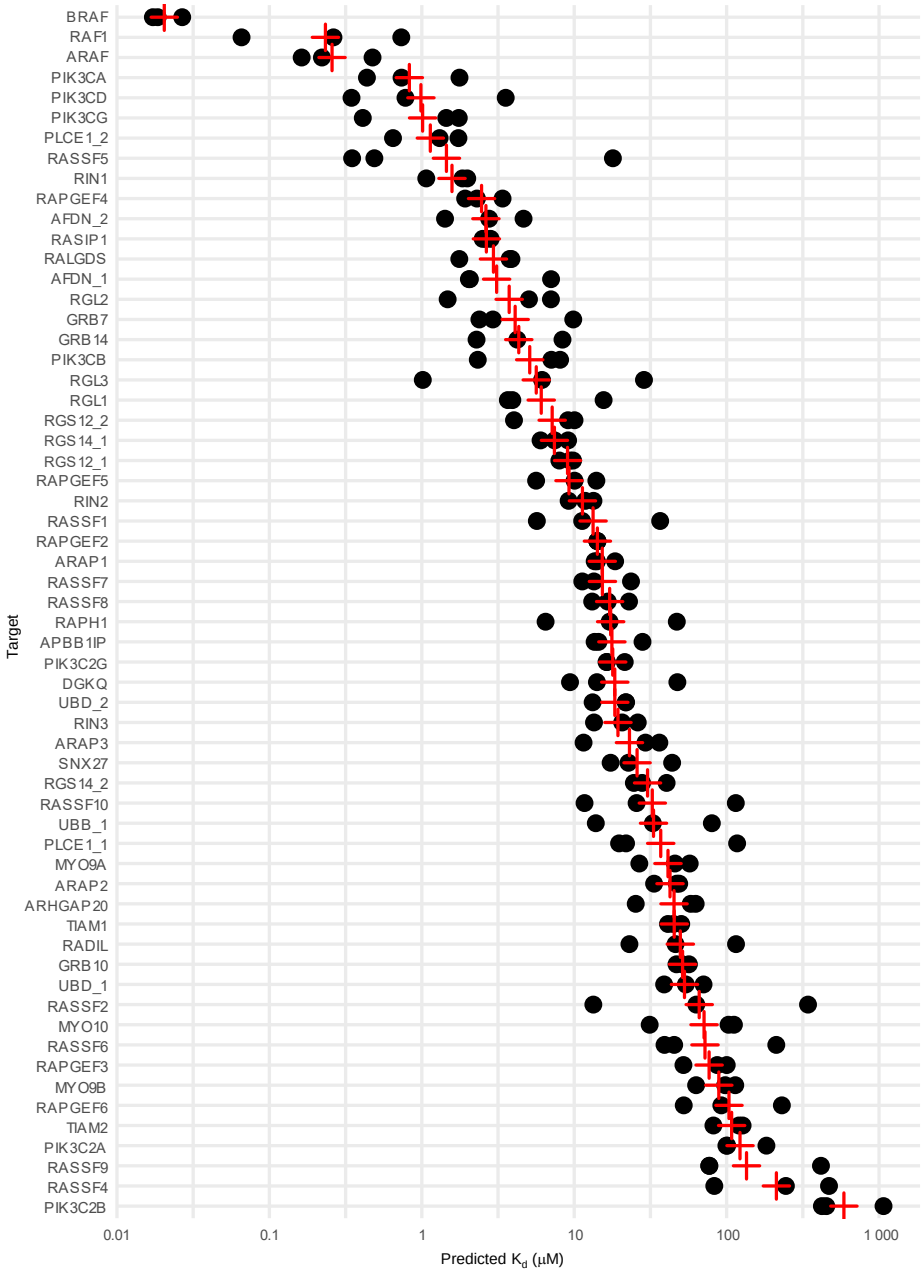

Figure 4.4 Results of the affinity prediction for all Ras-effector complex structures Visualization of predicted binding affinities. The three representative structures for each target are visualized as black dots, with the averaged affinity (based on averaged energy) is visualized as a red cross.

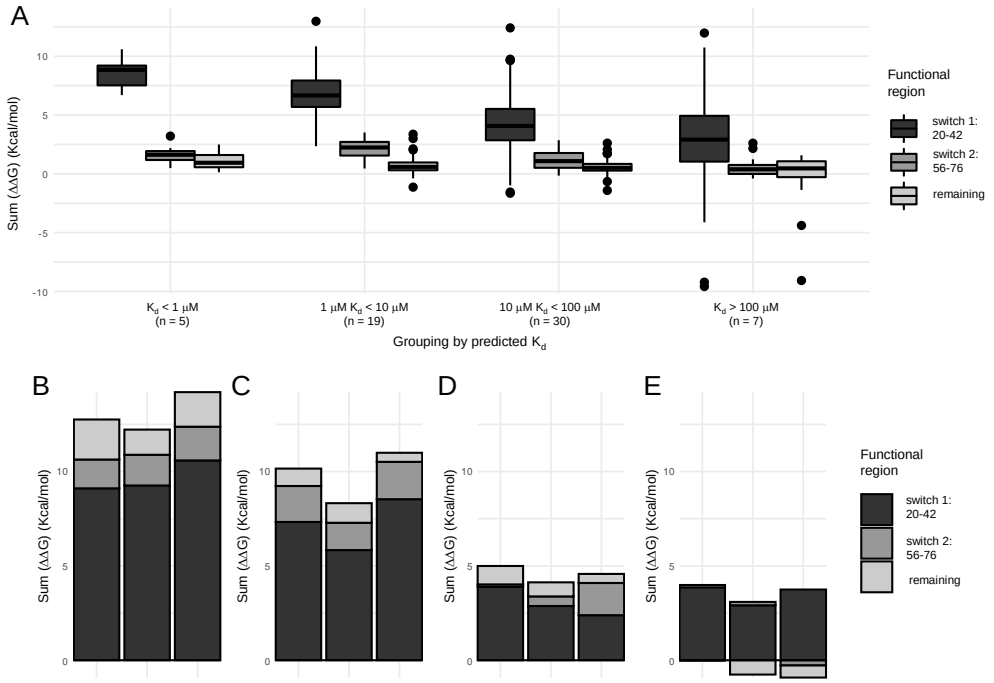

Figure 4.5 Switch contributions for the summed-up energy contributions grouped by their predicted binding affinity (A) Energy contributions were separately calculated for switch 1 switch or the rest of the Ras interface by summing up energy energies from the *in silico* alanine scan analysis. Complexes were grouped based on their predicted binding affinities into four groups. (B – E) Examples of the switch contributions for each of the four groups. Visualized are BRAF (panel B), AFDN_1 (panel C), ARHGAP20 (panel D), and PIK3C2B (panel E).

switch 2 as in the X-ray structures. For the two weak affinity binding groups (10 µM < K_d < 100 µM and (K_d > 100 µM) switch 1 contributions are in the range of 4 kcal mol^{-1} to 5 kcal mol^{-1} with small ($\sim$1 kcal mol^{-1}) contributions from switch 2 and remaining parts involved in interface formation. For the two strong affinity groups (K_d < 1 µM and 1 µM < K_d < 10 µM) switch 1 contributions increase to 6 kcal mol^{-1} to 9 kcal mol^{-1} with also increasing switch 2 contributions (1 kcal mol^{-1} to 2 kcal mol^{-1}). We also observed negative switch contributions (mainly for switch 1), indicating that these proteins are either not well predicted or non-binders. Indeed, all structures with negative switch energies are weak binders.

Branch pruning analysis using Ras-effector model structures

Next, we were interested in exploring surface mutations on Ras that would selectively influence the binding to some, but not all effectors. Both enhancing and inhibiting mutations are of interest. This could enable the engineering of the Ras effector system to respond to stimuli in different ways and to study selective sets of effectors. We previously reported a framework for the identification and evaluation of so called “branch pruning” mutations [115]. Since our protein is interacting with a many different effectors at the same time through the same interface it will be quite unlikely to identify mutations that selectively target only one protein. Instead, it is more likely that we will identify mutations that enhance in interaction with some proteins while inhibiting some others. Figure 4.6 shows a heatmap of all identified mutations of interest and their effect on all effectors. Some interesting mutations to highlight are mutations around I36, that are almost exclusively unfavorable while affecting almost every structure. D37 mutation are more selective and also exclusively disruptive. D38 is mixed, as our analysis of the hotspot already indicates. This is probably the best point to disrupt the system. Y40, interestingly, while being a ubiquitous hotspot, is not a good spot for engineering the interface because mutations seem to affect protein stability (compare with Figure 4.S8A). Also of interest is that we detect both mutations that increase and disrupt binding (Figure 4.S8B).

Assessment of rewiring of Ras-effector interaction on a systems level

Finally, we want to evaluate the behavior of the Ras effector system based on our predicted binding energies and based on the introduction of different branch pruning mutations. For this means, we went back to our mathematical model of the Ras-effector system that incorporated affinities and high-quality proteomics data in 29 human tissues [15]. Here, all 29 tissue systems were simulated at 20 % and 90 % GTP load on Ras. We are using exclusively the predicted binding energies for this (except RASSF8, see methods). Overall, the results for the systems without a branch pruning mutation are comparable to previous findings (Figure 4.7A; [15]). The Raf family members ARAF, BRAF

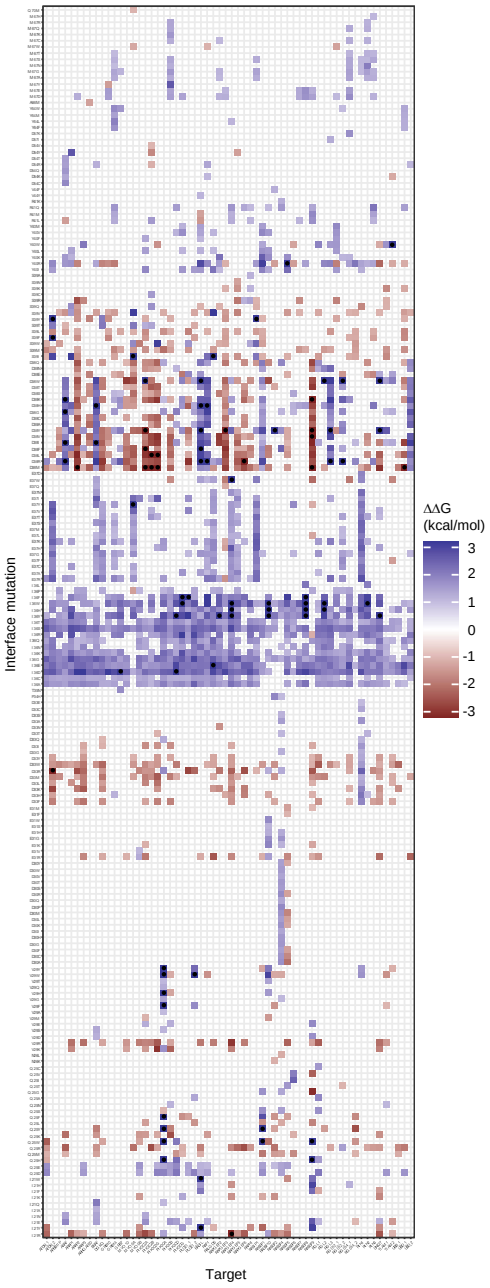

Figure 4.6 Energetic characterization of RAS interface mutations that affect effector binding The color scale is confined to the limits [–3.2 kcal mol^{-1}, 3.2 kcal mol^{-1}]. Hotspot residues with a ΔΔG $\geq$1 kcal mol^{-1} or $\leq$–1 kcal mol^{-1} were marked.

and RAF1 dominate the binding profile in complex with Ras. Other effectors that are in high amount in complex with Ras in at least one of the 29 tissues are RGL2, RASSF7, RASSF5, RASIP1, RALGDS, PIK3CD, PIK3CA, and AFDN.

Expanding on this, we introduced our branch pruning mutations to the mathematical model. In total, there are 200 interface mutations that do not significantly affect overall protein stability but affect binding to at least one of the effectors. Some of the branch pruning mutations are able to dramatically change the system in all tissues, as can be seen from the example of D38A (Figure 4.7B). With a single interface mutation, almost all RAF binding is quenched, and other effectors start to compete for the binding. Interestingly, which effectors come up depends on the tissue.

Next, we wanted to explore whether there are recurring states that the modelled system assumes and whether these states are dependent on Ras-GTP load, the tissue, or interface mutation. To this end, we applied uniform manifold approximation and projection (UMAP) to our systems to transform a high-dimensional space of absolute and relative effector binding into a 2D space. Then, we used OPTICS to identify areas of high density in this 2D plane and assigned them into 19 distinct clusters (Figure 4.7C). For each cluster, we picked three of the systems at random and visualized their relative effector binding (Figure 4.S9). Each of these clusters belongs to a different "state" that the Ras-effector system can be rewired into, with systems belonging to a specific state showing similar trends in effector binding. Many of the systems are dominated by ARAF binding, as it would be expected, but even for these there are differences in the secondary effectors. To understand what the attributes of different "states" of the systems are, two of them were picked and investigated for Ras-GTP load, tissue composition, and interface mutation status (Figure 4.7DE). We find that there are different ways a distinct state of the system can come together: for the state analyzed in Figure 4.7D, we can see that it is composed of many different tissues, but only a handful of interface mutations, most of them D38 mutations. This indicates that this state can be reached from many different tissues by a specific, recurring set of mutations. In contrast, the state analyzed in Figure 4.7E is entirely composed of a single tissue (lymph node) with many different mutations, indicated that this state can

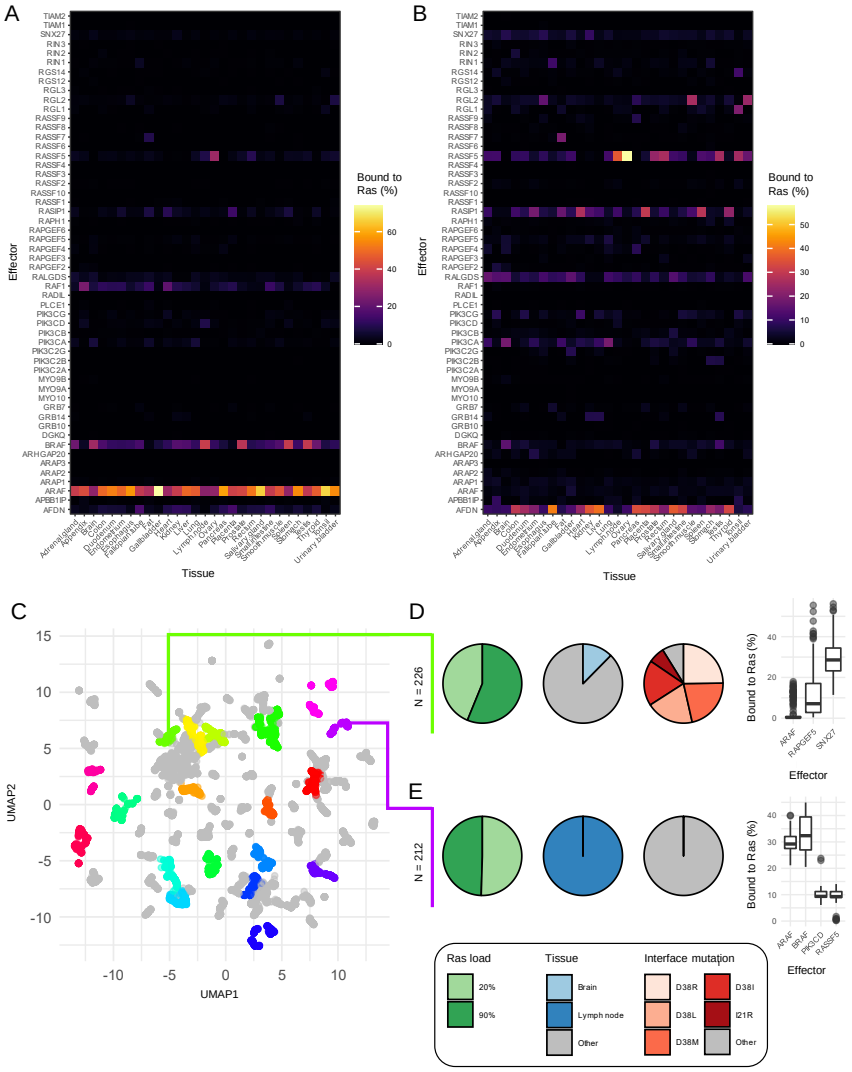

Figure 4.7 Ras-effector interaction rewiring at a systems level (A-B) Effects of interface mutation on the Ras-effector system. Heatmap of effectors bound to Ras at 20 % Ras-GTP load in 29 tissues in (panel A) WT interface and in (panel B) with the effects of a D38A mutation. (C-E) Overview over the possible states the Ras-effector system can assume. UMAP transformation of all simulated systems. Similar systems were identified using OPTICS clustering (panel C). 19 identified clusters are colored in rainbow colors, with systems classified as outliers colored in grey. Characterization of two identified state clusters in terms of Ras-GTP load, tissues, interface mutations, and most representative effectors (panels D and E). For the tissue and mutation pie charts, all groups that were smaller than 5 % of the total observations were collapsed into the “Other” group.

only be reached by a specific tissue. Interestingly, both states analyzed are diverse in terms of Ras-GTP load. To conclude, we identify 19 distinct states of the Ras-effector systems and show that there are different mechanisms on how these states are formed.

Finally, we wanted to explore to which extend, for a specific effector, it is possible to modulate its binding. For this, we visualized the possible changes to the relative amount of effector bound to Ras across all systems tested (Figure 4.S10). On the side of proteins that can be negatively influenced, mostly the high-affinity binders such as RAF family proteins show up, which is to be expected. Some effectors cannot be influenced by the branch pruning mutations, either because they are energetically not affected or because their concentrations in any of the 29 tissues do not leave them in a position to compete for binding. Examples for this would be RADIL or the TIAM family proteins. The effectors with the highest propensity to have their binding enhanced are AFDN, RADIL, SNX27, RASSF5. In summary, based on a large set of Ras-effector models, we predicted Ras branchegetic mutations and evaluate their binding in a tissue-specific Ras competition model. The Ras mutations, once introduced into cells or tissues, can be used as a tool to probe the contribution of specific effector pathways to an output or cellular phenotype.

4.2.4 Discussion

In this work we have shown a complete structural reconstruction of Ras and the RBDs of its effectors based on state-of-the-art technology. We used these structural models to investigate the workings of the interface between Ras and its effectors, as well as to search for and identify potential branch pruning mutations on Ras that would alter the behavior of the underlying system. We analyzed the effects on the steady state of the Ras effector system.

Recent advances in structural modelling, mostly by the development and release of AlphaFold2 and similar algorithms have pushed the field of structural bioinformatics forward. This development was crucial for the quality of this study. While a normal homology modelling pipeline would have been successful, the inclusion of both aspects of AlphaFold2 models was essential for the quality of

the results. On the one hand, finding additional potential complex templates diversified the possible configurations of the interface that we could use to generate our models. On the other hand, having high quality templates of the RBDs enabled us to improve the structural predictions. Interestingly, with all the advancements that AlphaFold2 brought, this combined AlphaFold2/homology modelling approach yielded more consistent and better results for this particular question.

There is a growing body of literature about how Ras dimerizes or forms multimers, or interacts with the membrane to modulate the signaling. All these aspects have been deliberately left out for this approach. The essence of the interaction of Ras and its effectors is the binary protein-protein interaction between the Ras switch regions and the RBD. This common feature was the focus of this work, and we believe that other factors such as dimer/multimerization of Ras, the composition of the membrane, etc., are only modulators for this interaction.

By creating a complete structural system, we were able to investigate and understand the interactions of Ras with its effector molecules on a different level. Crucially, it enabled us to analyze how *in silico* mutations of the system could affect its behavior. This is an interesting approach for a lot of different systems, not just the Ras effector system. However, there are certainly challenges. The prediction and verification of a protein-protein interface can be very complicated, and the techniques for modelling protein-protein interactions are not sufficiently developed to easily translate that approach to a larger scale. Probably the main reason why it was possible to construct this structural system for the Ras-effector system was because it is a conserved domain-domain interaction between homologue proteins. Although the sequence identity of the RBD sequences is not well preserved anymore, the structural fold of these domains has been preserved. Additionally, there is also the preserved mode of binding by the formation of the intermolecular β-sheet. These factors were favorable for the construction of the structural system and would need to be addressed if this approach were to be taken to a higher scale. Recent publications are already able to work on system-wide structural prediction for interactions [118]. These approaches are very promising for the structural characterization of known

protein complexes and can identify high-confidence novel interactions as well. However, for the exhaustive characterization of a full system, especially with transient interactions, the distinction between true and false positives seems to remain challenging.

Finally, we hope that our structural and systems analysis of the branch pruning interface mutations will enable interesting experimental setups that study different downstream pathways from Ras. Some of the more promising candidates are AFDN, RADIL, SNX27, RASSF5. The downstream pathways of Ras that are best characterized are the RAF-MEK-ERK signaling pathway and the PI3K-AKT signaling pathway. However, some of the other proteins might play a role in a physiological or pathophysiological context as well. AFDN is essential for the organization of adhesion between cells[119], function that is often impaired in cancer [120]. RADIL is also linked to cell adhesion, and recent data showed that knockdown of was linked to decreased cell proliferation and invasion [121]. SNX27 is also part of signaling pathways that link to cell adhesion and barrier function [122]. RASSF5 is a tumor suppressor and has been shown to inhibit growth and invasion and to induce apoptosis [123]. Importantly, all branchegetic mutations were studied on a systems level using a tissues specific Ras competition model. These models can easily be adapted for specific cell systems of interest, provided that estimates for Ras and effector abundances are available. Altogether, this work contributed to a quantitative understanding of a key cellular hub protein – Ras.

4.2.5 STAR Methods

Detailed methods are provided in the online version of this paper and include the following:

- Key resource table
- Resource availability

Lead contact

Further information and requests for resources should be directed to Philipp Junk (philipp.junk@ucdconnect.ie)

Code availability

The code will soon be deposited on Zenodo.

4.2.6 Method Details

Experimentally determined Ras-effector complex structures

Structures were downloaded from the PDB. In the case where multiple models were available, the best one by MolProbity score was selected [94, 95]. The PDB files were processed so that all GTP and Mg2+ annotations were in the expected format. The list of used template structures can be found in Table 4.S1.

Interface characterization

Hotspots residues were determined by FoldX *in silico* Alanine scan [74, 75, 23, 76, 77]. In detail, each residue on both RAS and the effector was mutated to alanine and the *in silico* change of binding energy ΔΔG was determined as such:

$$\Delta\Delta G_{alascan} = \Delta G(mut) - \Delta G(wt)$$

A positive ΔΔG value indicates that the mutated residue is involved favorably in the interaction, whereas a negative ΔΔG value indicates that the mutation to alanine improved the interaction between the two proteins. The standard error for ΔΔG values in FoldX is around ±0.8 kcal mol^{-1}. In order to identify the most important residues for the respective interaction, a ΔΔG cut-off of 1.6 kcal mol^{-1} was chosen for the investigation of crystal structures. Since the

energies are systematically lower in the modelled complex structures, a cut-off of 1.2 kcal mol^{-1} was chosen for those.

Based on the alanine scan results, the contribution of the two major functional regions in the RAS interface [124], switch 1 (residues 20-42) and switch 2 (residues 56-76) to the interaction were determined. For each functional region and the remainder of Ras, the ΔΔG values were filtered by abs(ΔΔG) >0.8 kcal mol^{-1} and then summed up. The definition of the switch regions used here is more generous than what is normally used in the literature, and it would be probably more correct to label them as "switch-influenced" regions. These residue ranges aim to capture the regions in the interface that are affected by the movement of switch 1 and 2 during the transition from the inactive GDP bound state to the active GTP bound state. The conserved intra-molecular β-sheet between the β2 sheet on Ras with the β2 sheet on the effector is evaluated by measuring the differences in inter-molecular distances between the β-sheets in the experimentally solved complex structures and comparing them to the ones in a structure of interest. This measurement has been named Measured Alignment Error (MAE) in this manuscript:

$$MAE\left(res_{eff}, res_{ras}\right) = \frac{\Sigma_{i=1}^{n}\sqrt{\left(dist_{model} - dist_{ref}\right)^2}}{n}$$

with $dist_{model}$ being the Euclidean distance between two residues in the model of interest; $dist_{ref}$ being the Euclidean distance between the residue on Ras and the corresponding residue in the reference structure, as determined by structural alignment using TMalign [125]; and n being the number of reference structures. The references used were the X-ray complex templates found in the PDB (Table 4.S1). When the MAE is calculated for X-ray complex templates, the comparison with itself is removed from the calculation.

AlphaFold2 determined single structures for RBDs

For each protein containing one or more potential RBD sequences as identified in [30], the full structure was obtained from the AlphaFold Protein Structure database [9, 59] (Release 1, accessed September 2021). The sequences of

the RBDs were obtained from Pfam [126] and UNIPROT [127]. The part of the structures that correspond to RBD sequences was subsequently extracted from the AlphaFold2 structures. The list of used templates as well as information about the extracted RBD domains can be found in Table 4.S2.

AlphaFold2 determined complex structures: generation and selection

AlphaFold2-multimer/ColabFold (version 1.3.0) was run on a local computer [128, 9, 129]. Multiple sequence alignments were generated from MS2seq [130, 131, 132, 107]. Complex models were generated with HRAS as interaction partner. For each target, five models were generated with additional template search and five without additional template search [128, 133]. The resulting model were relaxed as per ColabFold defaults [128, 134]. For each model, several metrices were evaluated. Firstly, AlphaFold2's Predicted Alignment Error (PAE) was taken into consideration. PAE is an estimation of the error of pairwise distances between residues, that can be used to assess how confident AlphaFold2 is in the inter-domain arrangement of its models. Secondly, FoldX [74, 75, 76, 77] interaction energies were determined for the complexes as described above. Finally, the expected orientation of the inter-molecular β-sheets was evaluated by MAE. The best model by MAE was selected for each target, and then all complex structures with a MAE > 1 Å were removed.

Homology modelling pipeline

Alignment generation, model generation and model evaluation were performed using the homelette homology modelling interface [116]. Inputs to the homology modelling pipeline were complex structures of Ras in complex with some effectors, either experimentally determined or selected from AlphaFold2 complex predictions, as well as AlphaFold2 models of all RBDs of interest. For each target, all combinations of the RBD template with all complex templates were used to generate 300 models of the target in complex with HRAS each. The different sources for the complex templates were run separately with slight differences in the modelling procedure. For complex templates of experimen-

tal origin, TMalign [125] was used to generate a structure-based sequence alignment based on the RBD in the complex template and the single RBD of the target structure. Then, with those two templates as inputs, models were generated. For complex templates of *in silico* origin, structure-based sequence alignments were generated with TMalign [125] as described. As an additional template a HRAS single structure (5P21 [109]) was used in the modelling process since the AlphaFold2-generated complex templates, in contrast to the complex templates of experimental origin, do not have information about the important heteroatoms GTP and MG2+ in the Ras part of the structure. Models were generated using modeller [89, 96] with the altMOD extension [97]. All models generated were evaluated using QMEAN [91, 92], MolProbity [94, 95], and SOAP [90] potentials. A combined score was determined based on borda count as such:

$$combined\ score(X) = \Sigma_{i=1}^{m} n - rank_i(X)$$

For an observation X with i...m being a collection of evaluation criteria and n the total number of observations. A structure with high borda score is a structure that performs well across all metrices.

For each of the different sources of complex templates (experimental or *in silico*), 300 models were selected in a first selection step based on the combined score. These 600 models per target were then further analyzed using FoldX. *In silico* interaction energies were determined and *in silico* alanine scan was performed (see Interface Characterization).

In addition to generating models for unknown targets, a set of validation models based on the experimentally solved models were generated as well. For each of the seven PDB complex structures, 300 models were generated. Inputs were restricted so that the structure to be modelled would not be used as a complex template, but only the remaining six experimentally derived complex templates. For each validation target, 300 models were selected as described above.

Using the results from the analysis with FoldX, representative structures were selected based on an unsupervised learning workflow. As the clustering

algorithm of choice, Ordering Points To Identify the Clustering Structure (OPTICS) [135] was used, as implemented in scikit-learn (`https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html`). OPTICS is a density-based clustering method, which unlike the more popular k-means clustering does not require a manually set input of the number of clusters. Additionally, OPTICS is able to label data points as outliers. Several approaches to feature selection and/or feature engineering, hyperparameters of OPTICS, as well as methods for selecting representative structures from clusters were evaluated against the set of validation models (Table 4.S3). To select the best combination of hyperparameters, for each combination, the *in silico* alanine scan results of the representative structures were correlated to the results of corresponding PDB structures, and the combination of hyperparameters with the most stable performance across all validation sets (by minimum *z*-score of the correlation against the PDB structure for all 7 validation sets) was chosen. The best combination of hyperparameters is highlighted in (Table 4.S3).

Estimation of binding energies

Supervised learning based on a number of FoldX-derived features was used in order to estimate binding energy of complex models. The features consisted of the FoldX interaction energy, the energy contribution of switch 1, 2 and the remainder of the Ras protein interface (see Interface characterization), and the $\Delta\Delta G$ values for hotspot residues on Ras (cut off 1.2 kcal mol^{-1}). All features were standardized and highly intercorrelated features were removed. Data preparation was performed in R.

The experimental measurements of the dissociation constant between Ras-effector complexes were collected from several publications (see Table 4.S5, also available as supplementary data). Effectors that were experimentally determined as non-binders were removed from the data set due to uncertainty how to encode them with the varying technical limitations on detectable binding energies at the time of their publication. Also, the models generated for RASSF3 seem to be outliers with regards to FoldX interaction energy (see Figure 4.7) and were therefore removed from the prediction. A test set was

manually chosen from the available experimental measurements to cover the full spectrum of experimentally determined interaction energies. At the end, this gave us a training set with 51 observations (17 × 3 models) and a testing set with 12 observations (4 × 3 models).

Supervised learning was performed in scikit-learn (`https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html`) using different combinations of feature selection algorithms and regressors. Feature selection was performed using either f-regression or mutual information as implemented by scikit-learn. Regression was evaluated for different algorithms with various hyperparameter spaces (see Table 4.S4). All combination of feature selection and regression were evaluated using Group-K-Fold cross validation within the training data, with k=5 and the groups corresponding to the three structural models chosen for each target. Models were evaluated using R2 score. The best performing combination was trained on the whole training data and evaluated against the test data. Finally, this model was used to predict the binding energies for the complex structures without experimental data.

Branch pruning

FoldX was used to evaluate the effect of interface mutation on the binding energies in the modelled complex structures. As previously described, [115] the FoldX command PSSM was used to evaluate the changes in binding energy, and the FoldX command PosScan was used to evaluate if mutations impacted the stability of RAS.

All mutations that were destabilizing HRAS in either of two structures (3TGP and 5P21) were removed from the analysis (cutoff 1.6 $\mathrm{kcal\,mol^{-1}}$). Then, between the three models for each target structures, it was checked if a mutation had a noticeable impact (>1 $\mathrm{kcal\,mol^{-1}}$ or <−1 $\mathrm{kcal\,mol^{-1}}$) in at least two of the three structures. If so, the changes in binding energy for all models above/below the cutoff were averaged.

Systems analysis

A mathematical model of the Ras-effector system was set up as previously described [30, 15]. The model is based on classic ligand-receptor kinetics according to the assumption of conservation of mass. A system of ordinary differential equations was set up and steady states were calculated as described. The reactions are expressed as such:

$$R + E_i \overset{(k_i, k_{-i})}{\longleftrightarrow} RE_i$$

With R representing the molar concentration of Ras, E_i the molar concentration of an effector of the set of $i \in (1, 2, ..., 54)$ effectors (all modelled proteins, except for proteins from the Ubiquitin family) and RE_i the molar concentration of a Ras-effector complexes. The complex is formed at rate k_i and dismantled at rate k_{-i}. These rates define the dissociation constant:

$$K_{d_i} = \frac{k_{-i}}{k_i}$$

Due to the assumption of mass conservation,

$$R_{tot} = R + \sum_{1}^{54} RE_i$$

and

$$E_{tot} = E_i + RE_i$$

the system to solve therefore is:

$$R_{tot} = R + \sum_{1}^{54} RE_i$$
$$E_{tot} = E_i + RE_i$$
$$K_{d_i} = \frac{(R * E_i)}{RE_i}$$

for the set of $i = \{1, 2, ..., 54\}$ and can be numerically solved for RE_i. The system was solved using SciPy.

The concentrations for the species in the model were taken from the supplementary data of [15], in which molar concentrations were derived from high-quality proteomics data set of 29 different human tissues [60]. The concentration of Ras proteins (HRAS, NRAS, KRAS) were pooled together and then multiplied with a loading factor to take into account the balance between active (GTP bound) and inactive (GDP bound) Ras. This loading factor was set to 0.2 for a normal RAS system, and 0.9 for a system hyperactivated by an oncogenic hotspot mutation.

The binding affinities were taken from the predicted binding affinities (see Estimation of binding affinities), with the exception of RASSF3 for which we are not confident in our structural predictions. An experimentally determined binding affinity for RASSF3 was substituted from [136]. Additional sets of binding affinities based on the branch pruning analysis were evaluated as well. For this, the predicted binding affinities were adapted by the ddG values from the branch pruning analysis. For the case that a system with hyperactivated Ras was considered, it was made sure that common oncogenic hotspots such as G12, G13 and Q61 do not influence the binding energies for any effector.

Some sets of parameters were not solvable in a physically meaningful solution space (negative concentrations) or at all, these parameter sets were removed from further analysis.

To gain an overview over all solved system, absolute and relative effector bindings were transformed using UMAP [137] and visualized. OPTICS

clustering [135] was performed on the UMAP transformed data (parameters: min_sample = 25, min_cluster_size=200).

Data analysis and visualization

All data analysis, unless otherwise noted was performed in R using the tidyverse environment (`https://www.R-project.org/`). Visualizations were generated using ggplot2 (H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.). Visualizations of protein structures were generated using PyMol (version 2.5.0) (urlhttps://pymol.org/2/support.html).

4.2.7 Supplementary Materials

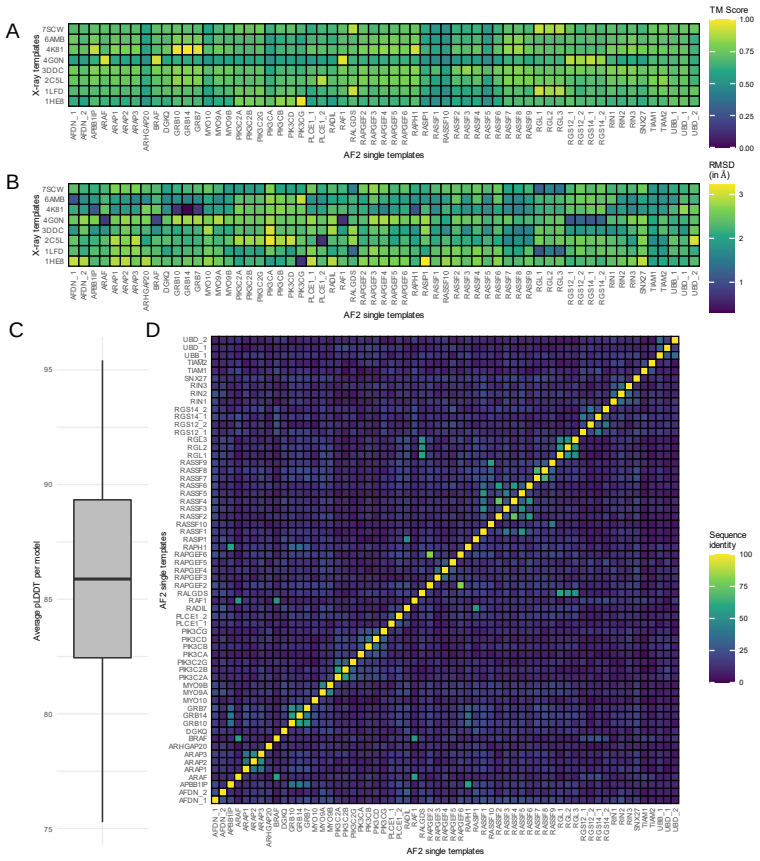

Figure 4.S1 Characterization and evaluation of AF2 RBD single templates retrieved from the AlphaFold Protein Structure Database (A) TM Score from the TMalign structural alignment program of AF2 generated RBD structures in comparison with the RBD domains of known complex structures. A TM Score between 0 and 0.3 corresponds to a random fold, whereas a TM Score of >0.5 indicates a similar fold. The TM score shown is normalised to the length of the AF2 template. (B) RMSD (in Å) of AF2 generated RBD structures compared to crystal structure references. RMSD was calculated using TMalign. (C) Distribution of average pLDDT values from the models retrieved from the AlphaFold Protein Structure Database. The pLDDT score is a local measure of how confidence estimated by AlphaFold2 when generating models. pLDDT scores >70 and >90 correspond to confident and very confident predictions, respectively. (D) Matrix of pairwise sequence similarity (in %) for all RBD/Ubiquitin-superfold domains investigated in this study.

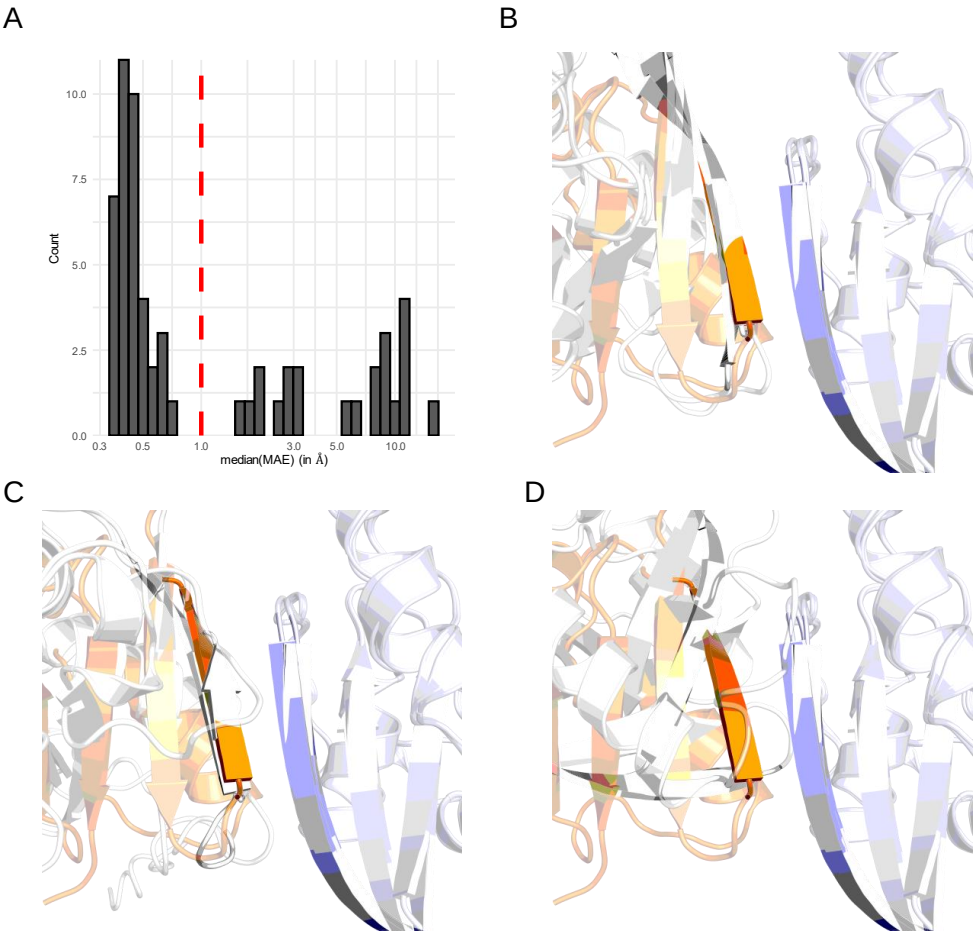

Figure 4.S2 Selection of AF2-generated complex templates by MAE (A) Histogram of the MAE distribution for the best structure for each target by MAE. The selection cut off was set to 1 Å. The x-axis is visualized in log10 transformation. (B) Visualization of b-sheet alignment for complex models generated with AlphaFold2 which have low median MAE value (GRB10: 0.35, RALGDS: 0.41). (C) Visualization of b-sheet alignment for complex models generated with AlphaFold2 which have moderate median MAE value (MY09A: 0.62, RASSF6: 0.68). (D) Visualization of b-sheet alignment for complex models generated with AlphaFold2 which have high median MAE value (RASSF7: 1.53, MYO10: 1.77). In B, C and D, Ras is displayed in light blue, RAF1 from the X-ray structure 4G0N is displayed in orange and the structural models generated with AlphaFold2 are displayed in white. The proteins were made semi-transparent except for β2 on RAS and β2 on the effectors.

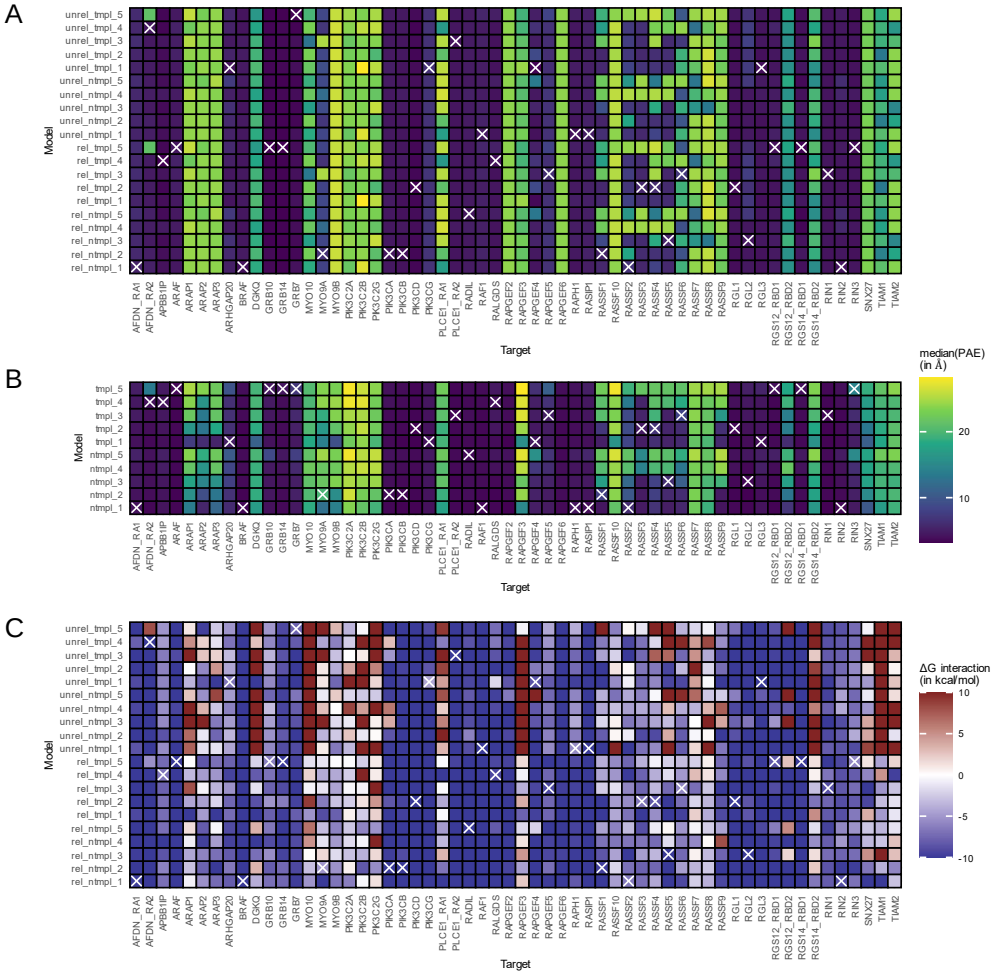

Figure 4.S3 Characterization of AF2-generated complex templates (A) Median Measured Alignment Error (β-sheet alignment) for AF2 determined complex structures in Angstroms. Selected structures are marked with a white cross. (B) Median Intra-molecular Predicted Alignment Error for all AF2 determined complex structures. For the calculation of the median, only PAE values between the two chains in the model were considered, not within the two chains. Selected structures are marked with a white cross. (C) FoldX interaction energy for all AF2 determined complex structures. Selected structures are marked with a white cross.

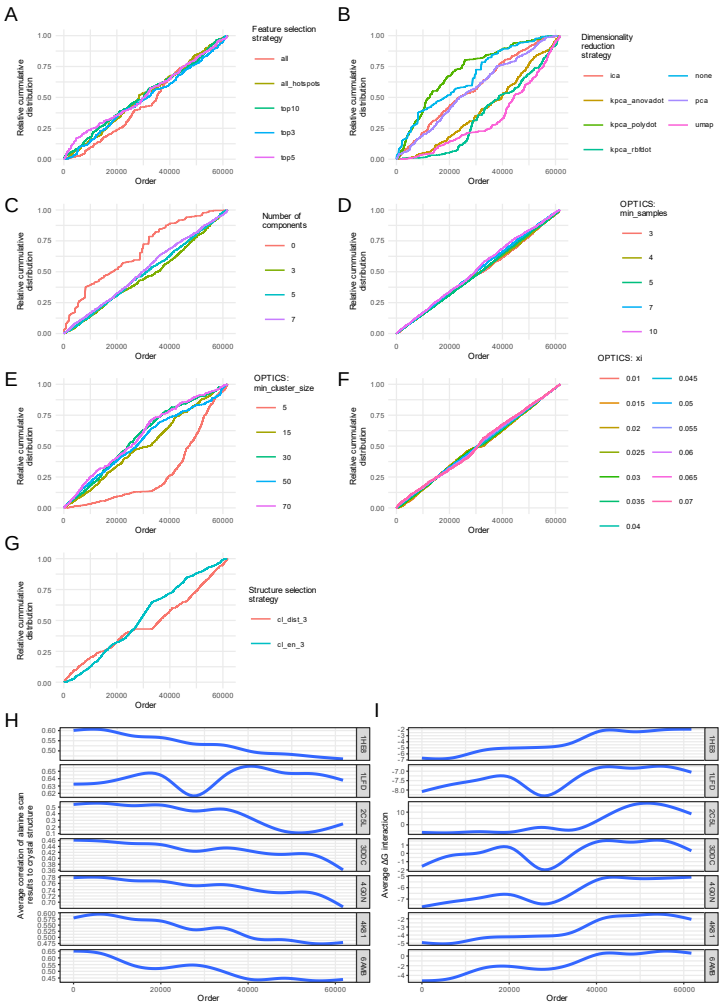

Figure 4.S4 Evaluation of hyperparameter space for clustering strategy to find good representative models based on FoldX alanine scan results in the validation set Relative cumulative distributions separated by hyperparameter in question (A-G). Order is established by sorting by average correlation of alanine scan results for the three selected models to the alanine scan results of the respective crystal structure. The hypermeter space consists of different feature selection strategies (A), dimensionality reduction methods (B), the number of selected components after a dimensionality reduction method (C), parameters for OPTICS min_samples (D), min_cluster_size (E), xi (F), and finally the approach for selecting the three representative structures from the cluster with the best average FoldX energy (G). Visualisation of average correlation to the crystal structure (H) and average FoldX interaction energy (I) for the cluster with the best average FoldX interaction energy across the ordering of combinations in the hyperparameter space.

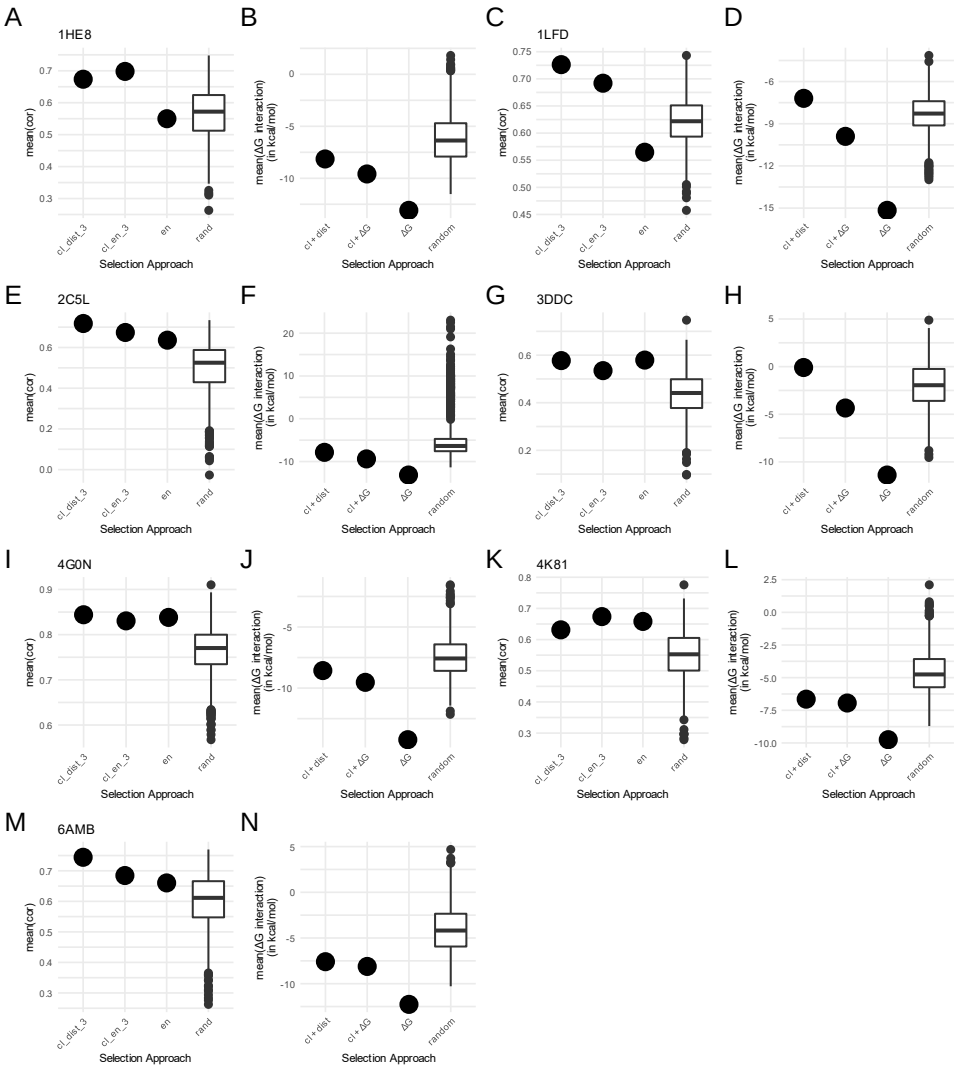

Figure 4.S5 Comparison of different strategies to select three representative structures for the validation set (A and B: 1HE8; C and D: 1LFD; E and F: 2C5L; G and H: 3DDC; I and J: 4G0N; K and L: 4K81; M and N: 6AMB). Visualization of the average correlation of the selection (A, C, E, G, I, K and M) and the average FoldX interaction energy (B, D, F, H, J, L and N). Selection strategy cl+dist is the selection of three representative structures after clustering from the best cluster by FoldX interaction energy based on lowest average distance in the cluster. Selection strategy cl+DG is the selection of three representative structures after clustering from the best cluster by FoldX interaction energy based on the best FoldX interaction energy. Selection strategy DG is the selection of three representative structures without clustering based on the best FoldX interaction energy. Multiple draws of three random structures are visualized as a comparison.

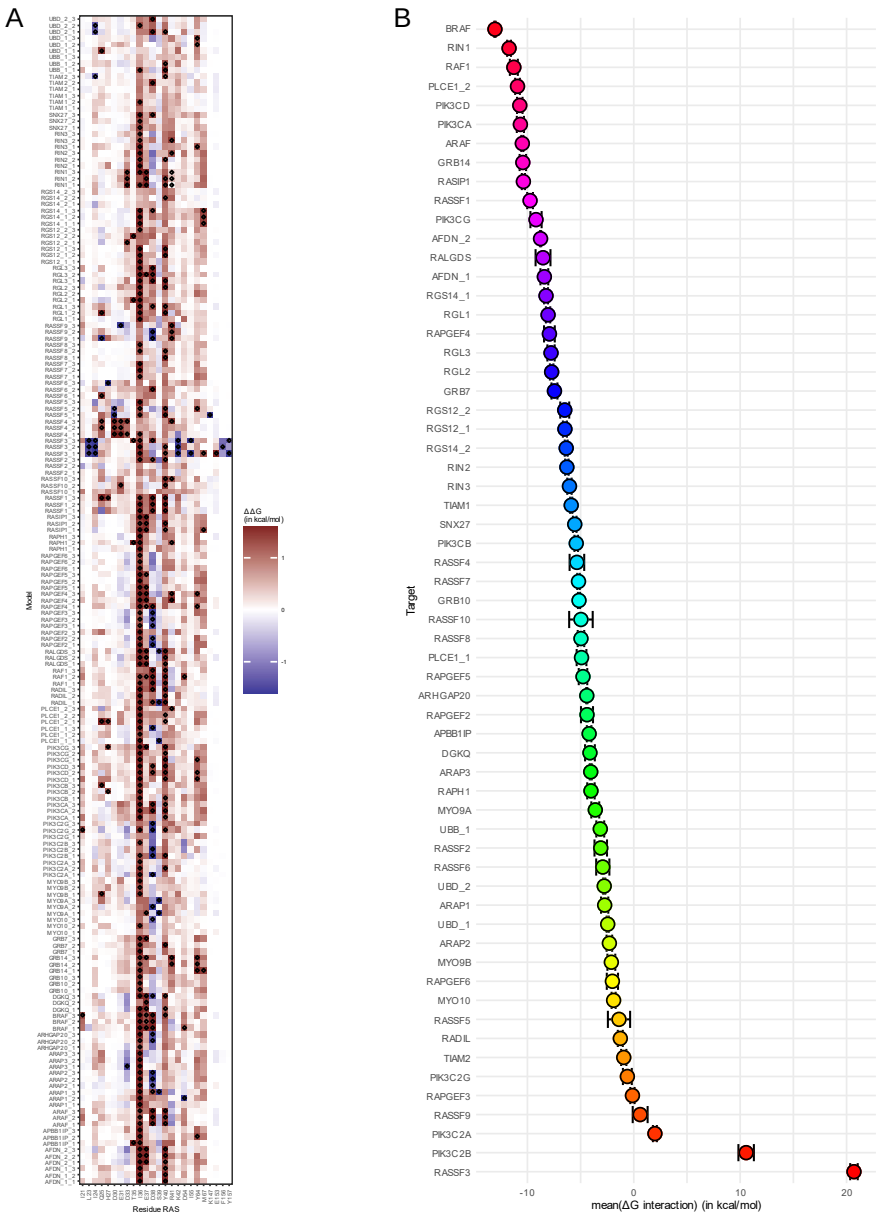

Figure 4.S6 FoldX energetic characterization of all models (A) Heatmap of FoldX alanine scan for all models. The color scale is confined to the limits [–1.6 kcal mol^{-1}, 1.6 kcal mol^{-1}]. Hotspot residues with a $\Delta\Delta G \geq 1.2$ kcal mol^{-1} or ≤ -1.2 kcal mol^{-1} were marked. (B) FoldX interaction energies average for the three representative structures on each target. The standard errors of the mean are displayed.

Figure 4.S7 Switch contributions of the three representative structures for each target For all modelled targets (A) – (BI), the individual switch contributions of the three representative models are visualized as bar charts.

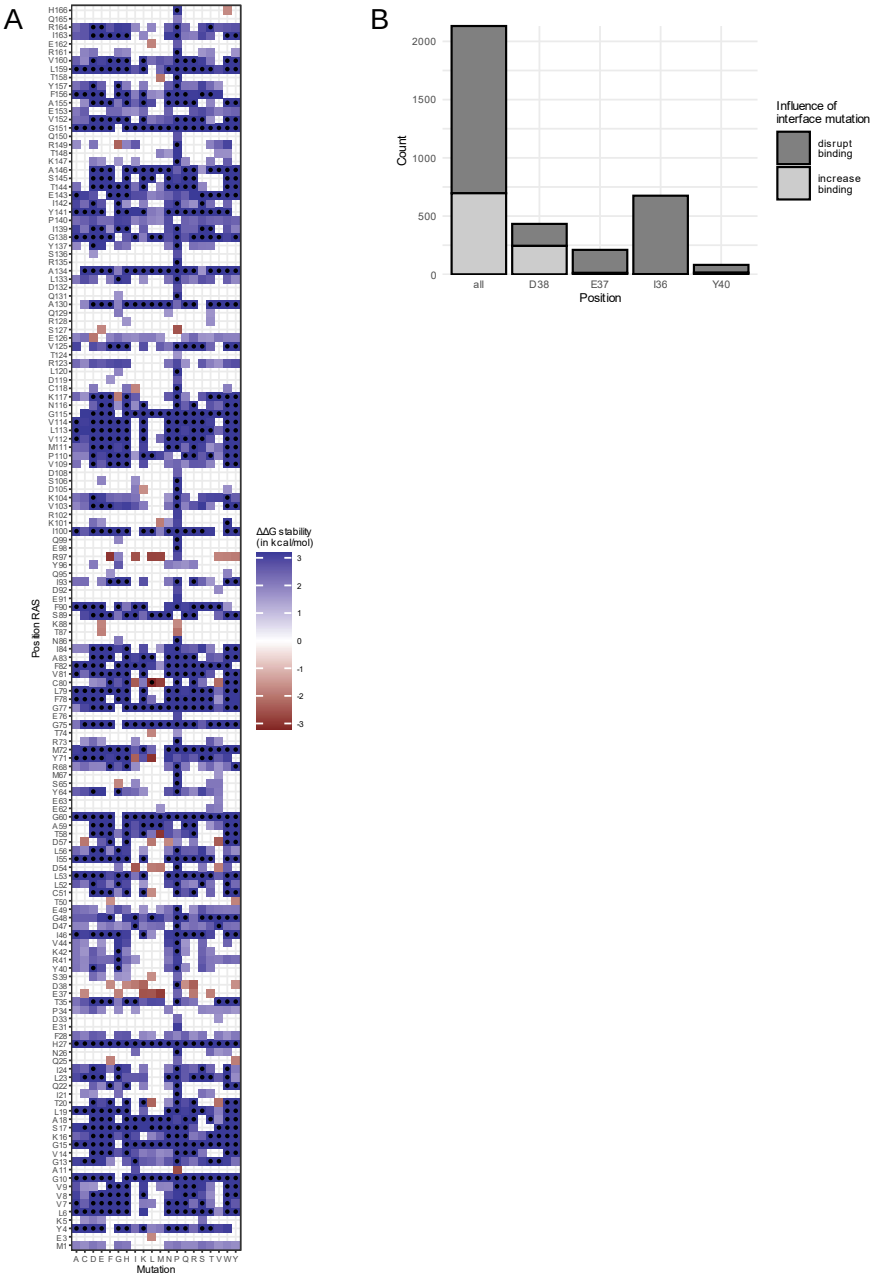

Figure 4.S8 A. Heatmap of FoldX stability of Ras proteins. The color scale is confined to the limits [−3.2 kcal mol^{-1}, 3.2 kcal mol^{-1}]. Hotspot residues with a $\Delta\Delta G \geq 1.6$ kcal mol^{-1} or ≤ -1.6 kcal mol^{-1} were marked. B. Count of branch pruning mutations and their effect on the whole analysis and different hotspots.

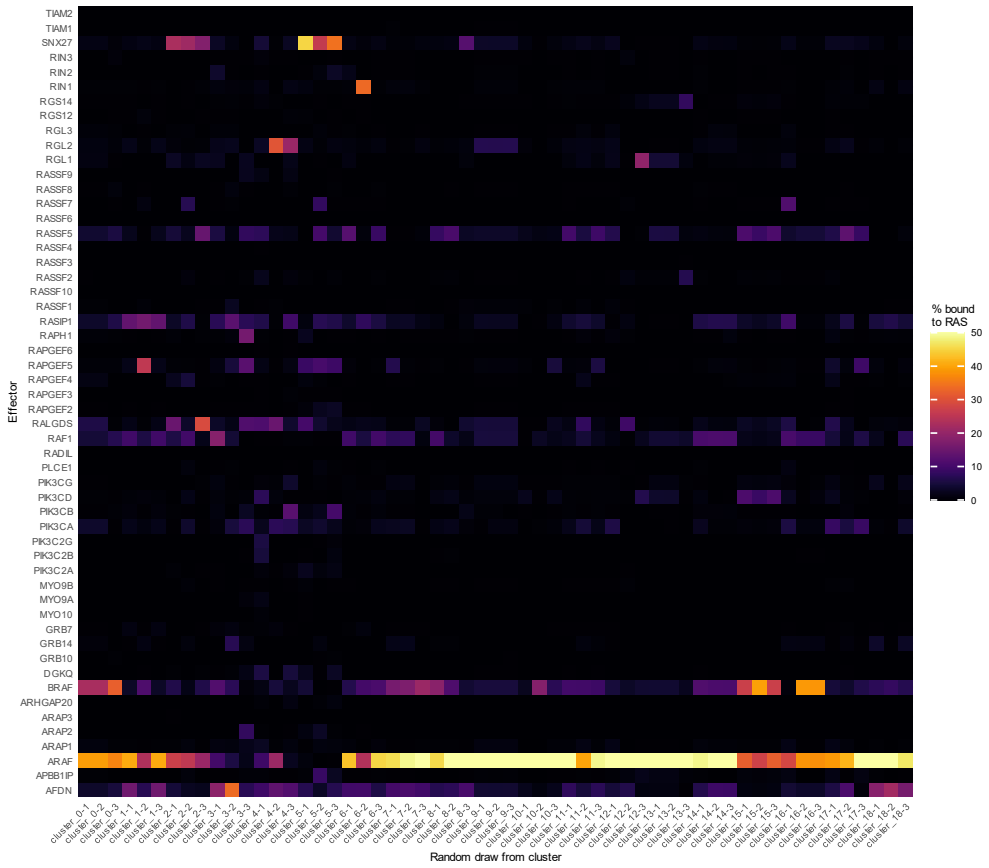

Figure 4.S9 Visualization of example system for the different UMAP-derived clusters. For each of the 19 clusters identified with OPTICS in the UMAP-transformed data, three members were randomly drawn, and the % of effectors bound to RAS was visualized. The colour scale restricted to [0 % - 50 %].

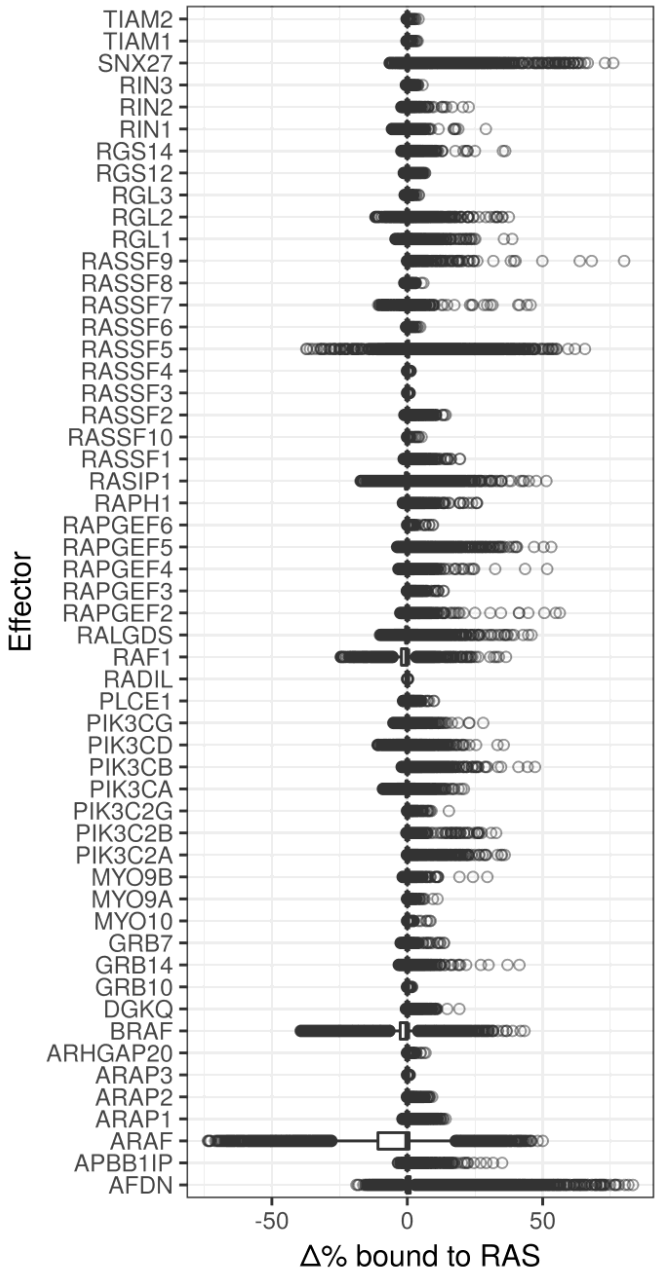

Figure 4.S10 Visualization of change in relative binding compared to WT system for each effector. Visualization for each effector how much the relative binding can change compared to the WT interface across all tested branch pruning mutations in all 29 tissues at 20 % or 90 % Ras-GTP load. Positive values indicate a relative increase in bound effector to Ras.

Table 4.S1 Complex templates.

Effector	Source	Identifier
PIK3CG	PDB	1HE8
RALGDS	PDB	1LFD
PLCE1	PDB	2C5L
RASSF5	PDB	3DDC
RAF1	PDB	4G0N
GRB14	PDB	4K81
AFDN	PDB	6AMB
RGL1	PDB	7SCW
AFDN_1	AlphaFold2 multimer	AF2c_AFDN_1
AFDN_2	AlphaFold2 multimer	AF2c_AFDN_2
APBB1IP	AlphaFold2 multimer	AF2c_APBB1IP
ARAF	AlphaFold2 multimer	AF2c_ARAF
ARHGAP20	AlphaFold2 multimer	AF2c_ARHGAP20
BRAF	AlphaFold2 multimer	AF2c_BRAF
GRB10	AlphaFold2 multimer	AF2c_GRB10
GRB14	AlphaFold2 multimer	AF2c_GRB14
GRB7	AlphaFold2 multimer	AF2c_GRB7
MYO9A	AlphaFold2 multimer	AF2c_MYO9A
PIK3CA	AlphaFold2 multimer	AF2c_PIK3CA
PIK3CB	AlphaFold2 multimer	AF2c_PIK3CB
PIK3CG	AlphaFold2 multimer	AF2c_PIK3CG
PIK3CD	AlphaFold2 multimer	AF2c_PIK3CD
PLCE1_2	AlphaFold2 multimer	AF2c_PLCE1_2
RADIL	AlphaFold2 multimer	AF2c_RADIL
RAF1	AlphaFold2 multimer	AF2c_RAF1
RALGDS	AlphaFold2 multimer	AF2c_RALGDS
RAPGEF4	AlphaFold2 multimer	AF2c_RAPGEF4
RAPGEF5	AlphaFold2 multimer	AF2c_RAPGEF5
RAPH1	AlphaFold2 multimer	AF2c_RAPH1
RASIP1	AlphaFold2 multimer	AF2c_RASIP1
RASSF1	AlphaFold2 multimer	AF2c_RASSF1
RASSF2	AlphaFold2 multimer	AF2c_RASSF2
RASSF3	AlphaFold2 multimer	AF2c_RASSF3
RASSF4	AlphaFold2 multimer	AF2c_RASSF4
RASSF5	AlphaFold2 multimer	AF2c_RASSF5
RASSF6	AlphaFold2 multimer	AF2c_RASSF6
RGL1	AlphaFold2 multimer	AF2c_RGL1
RGL2	AlphaFold2 multimer	AF2c_RGL2
RGL3	AlphaFold2 multimer	AF2c_RGL3
RGS12_1	AlphaFold2 multimer	AF2c_RGS12_1
RGS14_1	AlphaFold2 multimer	AF2c_RGS14_1
RIN1	AlphaFold2 multimer	AF2c_RIN1
RIN2	AlphaFold2 multimer	AF2c_RIN2
RIN3	AlphaFold2 multimer	AF2c_RIN3

Table 4.S2 Single templates. For each putative RBD, the AlphaFold2 prediction of the main UNIPROT sequence was downloaded from the AlphaFold Protein Structure Database (Release 1, accessed September 2021). The relevant domain was extracted either by using the domain information from PFAM, or for the proteins where PFAM had no RBD annotation, by using TMalign to identify the conserved ubiquitin superfold.

HGNC	UNIPROT ID	Residue start	Residue end
ARAF	P10398	19	91
BRAF	P15056	155	227
RAF1	P04049	56	131
PIK3CA	P42336	187	289
PIK3CB	P42338	194	285
PIK3CG	P48736	217	309
PIK3CD	O00329	187	278
PIK3C2A	O00443	421	509
PIK3C2B	O00750	375	463
PIK3C2G	O75747	285	371
RALGDS	Q12967	798	885
RGL1	Q9NZL6	648	735
RGL2	O15211	648	735
RGL3	Q3MIN7	613	700
AFDN_1	P55196_1	39	133
AFDN_	P55196_2	246	348
PLCE1_1	Q9P212_1	2012	2114
PLCE1_2	Q9P212_2	2135	2238
RIN1	Q13671	624	706
RIN2	Q8WYP3	787	878
RIN3	Q8TB24	877	963
SNX27	Q96L92	273	362
TIAM1	Q13009	765	832
TIAM2	Q8IVF5	810	881
ARHGAP20	Q9P2F6	194	295
ARAP1	Q96P48	1172	1261
ARAP2	Q8WZ64	1326	1420
ARAP3	Q8WWN8	1117	1210
DGKQ	P52824	395	494
RASSF1	Q9NS23	164	292
RASSF2	P50749	176	264
RASSF3	Q86WH2	79	186
RASSF4	Q9H2L5	174	262
RASSF5	Q8WWW0	236	364
RASSF6	Q6ZTQ3	218	306
RASSF7	Q02833	6	89
RASSF8	Q8NHQ8	1	82
RASSF9	O75901	25	119
RASSF10	A6NK89	4	133
RAPGEF2	Q9Y4G8	606	692
RAPGEF3	O95398	557	637
RAPGEF4	Q8WZA2	667	747
RAPGEF5	Q92565	240	321
RAPGEF6	Q8TEU7	749	835
RASIP1	Q5U651	144	259
RADIL	Q96JH8	61	164
APBB1IP	Q7Z5R6	176	263
RAPH1	Q70E73	269	355
MYO9A	B2RTY4	14	112
MYO9B	Q13459	15	114
MYO10	Q9HD67	1684	1791
RGS12_1	O14924_1	962	1032
RGS12_2	O14924_2	1034	1104
RGS14_1	O43566_1	302	373
RGS14_2	O43566_2	375	445
GRB7	Q14451	100	186
GRB10	Q13322	166	250
GRB14	Q14449	106	192
UBB_1	P0CG47_1	1	76
UBD_1	O15205_1	6	81
UBD_2	O15205_2	90	163

Table 4.S3 The hyperparameter space explored for the selection of representative structures. The best combination of hyperparameters is marked in bold text. For feature selection, hotspots were ranked by number of occurrences in all models for a specific target. min_samples, min_cluster_size and xi are hyperparameters of the OPTICS algorithm. For more information, please refer to the documentation: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html`

Hyperparameter	Options
Feature selection strategy	all, only hotspots, **top 10 hotspots + interaction**, top 5 hotspots + interaction, top 3 hotspots + interaction
Dimensionality reduction	**None**, PCA, ICA, UMAP, kPCA_rbfdot, kPCA_polydot, kPCA_anovadot
Number of components after dimensionality reduction	3, 5, 7
min_samples	3, 4, **5**, 7, 10
min_cluster_size	5, **15**, 30, 50, 70
xi	0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.04, 0.045, 0.05, 0.055, 0.06, 0.065, **0.07**
Selection of representative structures	Lowest energy cluster top 3 structures by FoldX interaction energy, **lowest energy cluster top 3 structures by smallest average distance**

Table 4.S4 The hyperparameter space explored for the regression of binding energies based on FoldX-derived features. The best combination of hyperparameters is marked in bold text. For all regressors, a two-step pipeline was used. In step 1, feature selection was performed using SelectKBest. In the step 2, the hyperparameter space for different regressors was explored. For more information, please refer to the documentation: `https://scikit-learn.org/stable`

Step	Model	Hyperparameter	Options
1	SelectKBest	k	1-27; best is **10**
		score_func	**mutual information regression**, f-regression
2	Lasso	alpha	0-1 in 1000 steps
	LassoLars	alpha	0-1 in 1000 steps
	Ridge	alpha	0-1 in 1000 steps
	ElasticNet	alpha	0-1 in 1000 steps
		l1_ratio	0.005, 0.01, 0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1
	Linear Regression		
	SVR	kernel	linear
		C	0.1, 1, 10, 100, 1000, 10000
		epsilon	0.001, 0.01, 0.1, 0.5, 1
	SVR	kernel	poly, **rbf**, sigmoid
		C	0.1, 1, 10, 100, **1000**, 10000
		epsilon	0.001, 0.01, 0.1, **0.5**, 1
		gamma	1 0.1, 0.01, **0.001**, 0.0001
	Random Forest Regressor	n_estimators	50, 75, 100, 125
		max_depth	None, 2, 3, 4, 5, 6
		max_features	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1
		bootstrap	True, False

Table 4.S5 Predicted K_d values in comparison with experimental K_d values. nb stands for not binding.

Target	Predicted K_d (μM)	Experimental K_d (μM)	PMID	Test/Training?
ARAF	0.257664	0.07	24441586	Training
BRAF	0.020433	0.04	24441586	Training
RAF1	0.233102	0.08	11292826	Test
PIK3CA	0.828134			
PIK3CB	5.106858			
PIK3CG	1.012316	2.9	11136978	Training
PIK3CD	0.986412			
PIK3C2A	122.4118			
PIK3C2B	588.1618			
PIK3C2G	17.89329			
RALGDS	2.95131	1	11292826	Training
RGL1	6.076352	3.5	9753431	Test
RGL2	3.744253	0.09	9753431	Training
RGL3	5.635062			
AFDN_1	3.096808	3.03	11292826	Training
AFDN_2	2.633623			
PLCE1_1	36.95527			
PLCE1_2	1.137651	0.82	15826668	Test
RIN1	1.577532	0.88	15826668	Training
RIN2	11.32005	11	15826668	Test
RIN3	19.36871			
SNX27	25.84413			
TIAM1	45.29575			
TIAM2	107.8572			
ARHGAP20	45.16912			
ARAP1	15.27132			
ARAP2	42.52022			
ARAP3	23.03746			
DGKQ	18.45376			
RASSF1	13.29281	39	15826668	Training
RASSF2	66.08577	147	33930461	Training
RASSF3		500	33930461	
RASSF4	212.0512	193	33930461	Training
RASSF5	1.451177	0.21	24441586	Training
RASSF6	72.1179	91	33930461	Training
RASSF7	15.30952	140	33930461	Training
RASSF8	17.0793	nb	33930461	
RASSF9	135.0758	179	33930461	Training
RASSF10	32.53544	nb	33930461	
RAPGEF2	14.19324	33	15826668	Training
RAPGEF3	76.75141	nb	15826669	
RAPGEF4	2.465967			
RAPGEF5	9.242346			
RAPGEF6	103.6208			
RASIP1	2.652572	nb	15826668	
RADIL	49.68458			
APBB1IP	17.66283			
RAPH1	17.34223			
MYO9A	41.23323			
MYO9B	88.87451	nb	15826669	
MYO10	71.10372			
RGS12_1	9.043251			
RGS12_2	7.166583			
RGS14_1	7.415029	14	24441586	Training
RGS14_2	30.27894			
GRB7	4.097531			
GRB10	51.39415			
GRB14	4.328095	3.6	26423700	Training
UBB_1	33.13215	nb	16310215	
UBD_1	52.91385	nb	16310215	
UBD_2	18.4631	nb	16310215	

4.3 Discussion

In this paper, we were exploring a structural systems biology approach to model the competition for RAS by its effector proteins. To this end, we generated structural models for all RAS effector complexes using state of the art methodology. Next, we analysed the binding interfaces and built a machine learning model predicting binding affinities. Using these affinities, we modelled the equilibrium state of the competition system. Finally, we explored interface mutations, which effect they have on the binding to the effectors, and how these changes affect the competition for the binding interface in the mathematical model.

For modelling the activities of RAS in the competition model, I assumed the GTP load and therefore the active state for RAS to be 20 % and 90 % for WT and oncogenic mutations of RAS, respectively. In reality, the GTP load of RAS is determined by the rates of nucleotide exchange and GTP hydrolysis, both intrinsic and supported by GAPs and GEFs. For some mutants, these rates have been determined experimentally [41]. According to these results, different mutants have different mechanisms in which they are proportionally more in the active state. For example, G12V-mutant HRAS is resistant to GAP-mediated hydrolysis. In contrast, G13D-mutant HRAS is less sensitive to GAP-mediated hydrolysis than WT HRAS, but still somewhat sensitive. Instead, it has an increased intrinsic and extrinsic nucleotide exchange rate. The used fixed values of 20 % and 90 % are useful to simplify the mathematical modelling and have been previously used when modelling the system [15].

A major part of this work was the generation of the complex models. The building and refinement of this pipeline, as shown in Figure 4.2 has been a lengthy process. At the time this project was started, AlphaFold2 was not around yet, so the pipeline was built purely on homology modelling. Also, docking strategies using HADDOCK have been explored. Over the course of this project, it has become clear that homology modelling was more applicable to our problem than docking. Then, with the release of AlphaFold2, more template structures became available, which improved our predictions. We also tried to model all the complexes with AlphaFold2-multi, however this approach only works for some of the complexes as shown in Figures 4.S1 and

4.S2. While many ideas have been tried for this pipeline, there are certainly possibilities to improve upon it, especially now that new technology based on AlphaFold2 is being developed.

We have been using FoldX for the energetic analysis of the complex interfaces. FoldX has been developed and tested with protein structures obtained from X-ray crystallography, we are however using it on structures created from computational methods. There could be concerns about the applicability of the FoldX force field for predicted protein structures, however a recent study shows that FoldX can be used on structures modelled with AlphaFold2 [138]. This is in line with what we observe in this paper on the validation set of protein structures, where the energetic predictions between the experimental and predicted protein structures are well aligned. Nonetheless, it would have been interesting to confirm the energetic predictions with other methods such as the Rosetta force field [88].

While the advance of technology for the prediction of protein structures is impressive, it is important to keep in mind that these are computational models and should be validated by experimental methods. This is in my opinion the biggest limitation of this work, and it would be interesting to see validation experiments of the predictions presented in the paper.

For protein structures, the gold standard is, and probably will be for a long time, X-ray crystallography. However, computational method such as AlphaFold2 are extremely valuable tools, as the turnover time to generate a structure is so much quicker. In addition, AlphaFold2 estimates how well it thinks it is able to predict a protein structure, which can be used to evaluate carefully whether to trust a model or not.

We have compared the list of interface hotspots we have identified in this work against variants in RAS that are known to lead to malignancies. Recently, an interesting paper was published evaluating the potential for different point mutations in KRAS to lead to hyperactivation [139]. Comparing our set of system-rewiring mutations against their mutation only had a small overlap. Of the 97 Ras variants they were studying, there was an overlap of four mutations. None of these mutations had any major impact on the competition of effectors for the Ras binding interface. Similarly, we compared the rewiring interface

mutations against known variants of Rasopathies. There was no overlap. This indicates that in the case of KRAS, disease-causing mechanisms are not based on modulating the interaction interface, but, as it has been well known for a long time, based on modulating the active state of RAS. This is why the mutations we were interested in do not coincide with oncogenic or other malignant mutations, because those tend to be localized around the nucleotide binding site.

There are multiple directions in which this work can be expanded on. Firstly, additional components of the RAS system can be structurally and mathematically modelled. For this, GAPs and GEFs, the proteins that control RAS activity and which are also competitors for the same interface as the effectors, come to mind. While they do not share the identical binding mode to the effectors, there are crystal structures of them available on the PDB which could be used as the basis for the model generation. Similarly, important downstream modules of RAS signalling, such as the MAP kinase pathway for example, could be modelled in the same way. It would be very interesting to see how much of the known complexity of MAPK signalling could be recapitulated using parameters derived from structural models.

Additionally, it would be interesting to move from static protein structures to dynamic protein structures with the inclusion of molecular dynamic simulations. RAS in particular is a very dynamic protein, so the analysis of the complex models with MD simulations could potentially shed some more light on how the binding interface works. Oncogenic mutations of RAS have been shown to affect the dynamics of the protein, in particular in the switch regions which assemble the interface. Using MD simulations could be an interesting approach to investigate if oncogenic mutations affect different RAS effector complexes differently.

5 Analysis of context-specific KRAS-effector (sub) complexes in Caco-2 cells

5.1 Introduction

Proteins control cell fates by being organised into complex signalling networks. RAS is one of the signalling hubs in the cell, controlling cellular states such as proliferation, differentiation and apoptosis [30, 33]. RAS is commonly mutated in many cancer types such as pancreatic cancer, lung cancer and colon cancer. The functionality of RAS is mediated by interaction of active RAS with effector proteins, although many details of these signalling events are currently not well understood [29, 34].

In particular, it is unclear how exactly signalling outcomes are determined and which effectors play a role for the respective outcomes. In a previous study by a colleague from my group, it was hypothesized that only a small subset of the effector proteins is able to bind RAS consistently on their own, with a much larger group being possibly dependent on some form of recruitment to the plasma membrane based on a specific stimulus [15]. This could be a mechanism by which cells (or other extracellular signals) could control and modulate their signalling response.

Based on this hypothesis, we investigated the interactome of KRAS using AP/MS in Caco-2 cells, an early colorectal cancer cell line, after overexpressing RAS and oncogenic mutations of RAS and treating the cells with different stimuli and inhibitors. In particular, we were interested in potential changes of the interactome based on different oncogenic variants or treatment conditions.

5.2 Analysis of context-specific KRAS-effector (sub) complexes in Caco-2 cells

Camille Ternet[1,2,*], **Philipp Junk**[1,2,*], Thomas Sevrin[1,2,*], Simona Catozzi[1,2], Giorgio Oliviero[1], Kieran Wynne[1,3], Christina Kiel[4,1,2]

[1] Systems Biology Ireland, School of Medicine, University College Dublin, Dublin 4, Ireland

[2] UCD Charles Institute of Dermatology, School of Medicine, University College Dublin, Dublin 4, Ireland

[3] Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Dublin 4, Ireland

[4] Department of Molecular Medicine, University of Pavia, 27100 Pavia, Italy

[*] Equal author contribution

This manuscript was **published** by **Life Science Alliance** on the 09/03/2023. It is available at `https://doi.org/10.26508/lsa.202201670` under an Open Access license.
A preprint is available at `https://doi.org/10.1101/2022.08.15.503960`.

The content is identical to the version available on BioRXiv, with changes to formatting and minor changes to bibliography. The supplementary figures, which are currently not available on BioRXiv, were added as well.

5.2.1 Abstract

Ras is a key switch controlling cell behavior. In the GTP-bound form, Ras interacts with numerous effectors in a mutually exclusive manner, where individual Ras-effectors are likely part of larger cellular (sub)complexes. The molecular details of these (sub)complexes and their alteration in specific contexts is not understood. Focusing on KRAS, we performed affinity purification (AP)-mass spectrometry (MS) experiments of exogenous expressed FLAG-KRAS WT and three oncogenic mutants ("genetic contexts") in the human Caco-2 cell line, each exposed to 11 different culture media ("culture contexts") that mimic conditions relevant in the colon and colorectal cancer. We identified four effectors present in complex with KRAS in all genetic and growth contexts ("context-general effectors"). Seven effectors are found in KRAS complexes in only some contexts ("context-specific effectors"). Analyzing all interactors in complex with KRAS per condition, we find that the culture contexts had a larger impact on interaction rewiring than genetic contexts. We investigated how changes in the interactome impact functional outcomes and created a shiny app for interactive visualization. We validated some of the functional differences in metabolism and proliferation. Finally, we used networks to evaluate how KRAS effectors are involved in the modulation of functions by random walk analyses of effector-mediated (sub)complexes. Altogether, our work shows the impact of environmental contexts on network rewiring, which provides insights into tissue-specific signaling mechanisms. This may also explain why KRAS oncogenic mutants may be causing cancer only in specific tissues despite KRAS being expressed in most cells and tissues.

5.2.2 Introduction

The interactome of a cell, like a social network, refers to the entirety of interactions of cellular molecules, in particular protein-protein interactions (PPIs) [16]. These interactions form a network and impact the spatial protein localization and functional organization of a cell. Networks adapt to internal and external cues by converting the signal in responses to stimuli into a plethora of possible

output functions that drive cell fates and phenotypes. PPIs, as the core of signaling networks, impact how signals are transduced, and alterations in cellular networks are often linked to diseases, particularly complex diseases such as cancer [16, 140]. Mutations in oncogenes can perturb PPI networks [141] when protein catalytic and binding functions are affected resulting in alterations in the proteins' binding interfaces [23].

The oncoprotein KRAS is an example of a hub signaling protein, as it is part of a highly interconnected and dynamic network capable of interacting with many other proteins [29]. Oncogenic mutations in KRAS rewire interactions and signaling pathways [142]. KRAS belongs to the Ras superfamily of GTPases and acts as a molecular switch that cycles between an inactive guanosine diphosphate (GDP)-bound state and an active guanosine triphosphate (GTP)-bound state. The GTP-bound Ras protein mediates binding to several downstream proteins, thereby controlling essential and diverse cellular processes such as survival, polarization, proliferation, differentiation, apoptosis, and migration [33, 30]. It is still enigmatic how Ras does all of it. However, what is known is that a class of proteins, called "effectors", plays a critical role [29, 110, 112].

Ras effectors are defined as proteins that bind much stronger (i.e. with higher affinity or lower K_d-value) to Ras-GTP than to Ras-GDP. Their interaction with Ras-GTP relies on a domain with a ubiquitin-like topology of three types: the Ras-binding domain (RBD), the Ras association (RA) domain, or the PI3K_rbd, which will herein collectively be referred to as RBDs. All effector RBDs recognize the same switch regions of Ras-GTP, which results in mutually exclusive binding [30, 143]. While the presence of an RBD is a necessary condition to qualify as an effector for Ras-GTP, it is not a sufficient criterion. Indeed, for a total of 56 effectors that contain RBD domains, the binding affinities between Ras-GTP-effector complexes are known (either from experiments or computational predictions) to vary, and some are predicted not to bind at all (Fig. 5.S1) [29, 30, 136, 144, 145].

In addition to affinities between the RBDs and Ras-GTP, protein abundances are important for complex formation. In a previous study we used protein abundances together with binding affinities in a mathematical model to predict

the amount of each of the 56 effectors in complex with Ras-GTP in 29 human tissues [15]. Surprisingly, only nine effectors form significant complexes ($\geq$5%) with Ras-GTP in at least one of the 29 tissues (here referred to as group 1 effectors). These results let us wonder about the relevance of the remaining effectors, some of which are well-established effectors, such as PI3-kinase (PI3K) [146]. As effectors are generally multi-domain proteins, we reasoned that domains that can transfer effectors to the plasma membrane (PM), where Ras-GTP is localized, can increase the number of complexes formed between Ras-GTP and effectors [30]. Indeed, seminal work by Kholodenko and colleagues has demonstrated that membrane anchoring of both interacting proteins strongly increases the average lifetime of complexes, i.e. the "piggyback" mechanism [111]. Indeed, when we applied the piggyback mechanism to the Ras-effector model, we identified 32 effectors that are predicted to form significant complexes with Ras-GTP only with an additional domain recruited to the PM (here referred to as group 2 effectors) (Fig. 5.S1). These effectors were predicted to be recruited to the PM in response to specific conditions (e.g. inputs/stimuli/growth factors) [29, 30]. The remaining 15 effectors are never predicted to be in significant complex with Ras-GTP and are likely no true Ras effectors (here referred to as group 3 effectors).

Colorectal cancer (CRC) is the fourth leading cause of cancer death worldwide. A statistic update reports that there were 1.8 million new cases worldwide in 2018, with a significant burden shift from the old to the younger individuals [147]. CRC develops through a complex sequence of processes involving an accumulation of epigenetic and genetic alterations, where one of the major driver mutations appears to be KRAS mutations and specific pathways that regulate cell growth and differentiation [148]. The most frequent KRAS mutations found in CRC are single-point mutations found at codon 12 with the G12D, G12V, and G12C mutations, often followed by codon G13 and Q61 [27, 149]. Oncogenic KRAS leads to an accumulation of constitutively active KRAS mutated proteins leading toward the activation of diversified downstream signaling pathways such as the RAS/RAF/MEK/ERK signaling pathway and the PI3K/AKT signaling pathways which were extensively studied in RAS effectors

cancer context [150]. However, there is evidence that other RAS effectors play a role in cancer [151].

In this work, we experimentally probed context-specific network rewiring of KRAS exogenously expressed with a FLAG-tag in immortalized human Caco-2 cells. This cell line, derived from human colorectal adenocarcinoma cells, harbors somatic APC mutations and CTNNB1 (i.e., β-catenin) mutations, but is wildtype (WT) for KRAS [61]. To probe different "genetic contexts", we exogenously expressed KRAS WT and three oncogenic mutations frequently found in CRC (G12V, G12D and G12C) to probe different "genetic contexts". To probe different "culture contexts" we grew Caco-2 cells in various growth media mimicking tumor microenvironments (TME) that are known to impact CRC maintenance, progression, and metastasis, which have been described earlier in connection with oncogenic KRAS. Interleukin-6 (IL-6) and tumor necrosis factor-alpha (TNF-α), both part of the inflammatory response found in the TME, are factors of those growth culture contexts [152, 153, 154]. (Patho)physiological conditions such as hypoxia [155, 156] (mimicked by Dimethyloxalylglycine, DMOG, [157]), epidermal growth factor (EGF), and prostaglandin E2 (PGE2) also play a role in CRC and KRAS TME [158, 159, 160] and were selected as growth conditions here. Each combination of genetic and culture contexts was analyzed separately in affinity purification-mass spectrometry (AP-MS) experiments [161, 162] to determine KRAS-mediated complexes. Our study provides an in-depth reconstruction of PPI networks mediated by oncogenic KRAS-effector proteins in specific relevant (patho)physiological colon contexts. Additionally, by identifying different levels of network organization called subcomplexes, we further detailed the downstream pathways mediated by effectors of KRAS and linked them to functional outputs.

5.2.3 Results

Analysis of KRAS-mediated networks in different genetic and culture contexts

We conducted AP-MS experiments to characterize the PPI landscape of both the WT and oncogenic mutant forms of KRAS in different growth conditions. KRAS protein variants were exogenously expressed as FLAG-tagged proteins under the control of a doxycycline-inducible promoter [163]. As observed earlier [163], the promotor shows some leakiness even without doxycycline. As we aimed to express FLAG-KRAS at relatively physiological levels, doxycycline was only added to express the FLAG-KRAS WT proteins at a dose that resulted in equal expression levels compared to the FLAG-KRAS mutant proteins expressed without doxycycline (Fig. 5.S2).

To analyze the KRAS WT and mutant interactomes in different growth media ("culture contexts") that mimic conditions relevant in the colon and CRC, Caco-2 cells were grown 4 h after transfection for 24 h in minimal medium (DMEM with 2 mM L-glutamine) supplemented with either IL-6, TNF-α, PGE2, epidermal growth factor (EGF) or the HIF-hydroxylase inhibitor DMOG at different concentrations (20 and 200 ng mL^{-1}) before the AP-MS experiment was conducted. Altogether, we tested four "genetic contexts" (FLAG-KRAS WT, G12V, G12D and G12C) and 11 "culture contexts" (minimal medium, and each two concentrations of IL6, TNF-α, PGE2, EGF, and DMOG in minimal medium), resulting in 44 condition-specific AP-MS experiments (Fig. 5.1).

To identify high-confidence interacting proteins for each condition, the label-free quantification (LFQ) intensities data for all proteins were filtered in a series of steps (see Materials and methods). Specifically, data for each MS run (44 $\times$ three biological with two technical replicates = 264) was visualized as histogram to filter out 8 runs with very few proteins identified (Fig. 5.S3). Further, for a protein to qualify for the high-confidence list, it had to be detected in at least 60 % of the technical and biological replicates for a specific condition. In addition, only proteins that were significantly enriched compared to the beads only control were included. Technical replicates were merged using the median

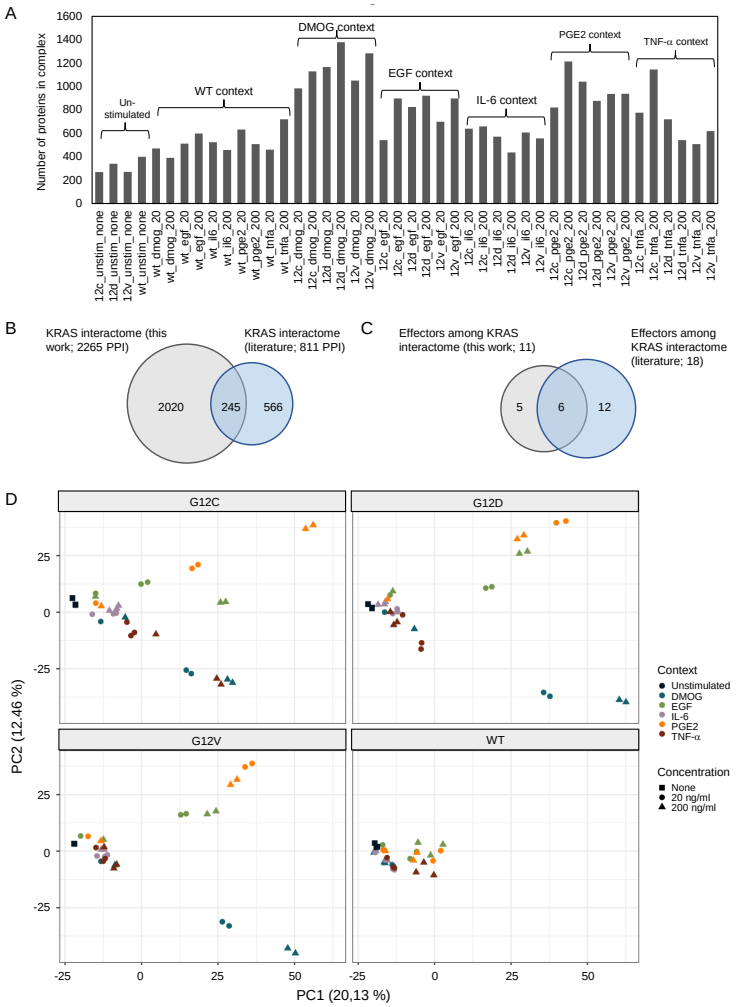

Figure 5.1 Global analysis of context-specific KRAS WT and mutant interactomes. (A) Number of proteins identified in KRAS WT and mutant AP-MS experiments after filtering. The conditions are unstimulated (minimal medium), DMOG (20 and 200 ng mL^{-1} in minimal medium), EGF (20 and 200 ng mL^{-1} in minimal medium), IL6 (20 and 200 ng mL^{-1} in minimal medium), PGE2 (20 and 200 ng mL^{-1} in minimal medium), and TNF-α (20 and 200 ng mL^{-1} in minimal medium). (B) Overlap of all interactions identified in at least one condition with the literature KRAS interactome described in [29]. (C) Overlap of the same datasets as in panel B but focusing on effector proteins. (D) Principal component (PC) analysis performed on label-free quantification intensity (LFQi) and executed with MS log2-transformed data after filtering on the whole AP-MS dataset. Colors indicate the different growth conditions, i.e., DMOG, EGF, IL-6, PGE2, TNF-a and unstimulated (unstim), and shapes indicate the concentration of the conditions (none, 20 ng mL^{-1} and 200 ng mL^{-1}).

LFQ intensity. To verify the robustness and applicability of the protocol, we analyzed KRAS expression levels in the complete dataset. The LFQ intensity of the KRAS bait is comparable across all AP-MS conditions (Fig. 5.S4), which suggests that a similar quantity of FLAG-KRAS proteins binds to the magnetic beads across all APs.

A total of 2265 high-confidence PPIs were identified in the 44 contexts, with an average of 725 PPIs per condition (Fig. 5.1). Although KRAS is a small protein, a large number of interaction partners is not too surprising, as many of those are expected to be not direct binary physical interactors but rather bind via third proteins (i.e. effectors or other proteins that enable compatible complex formations). Of note, less interactors are generally found for conditions in minimal medium (cf. unstimulated conditions in Fig. 5.1A), supporting our initial hypothesis that microenvironmental contexts play a significant role in KRAS complex formations. Further, the FLAG-KRAS WT APs generally have less PPIs, which can likely be explained by the fact that effectors bind KRAS predominantly in its GTP-bound form (in fact, no effectors are detected in any of the FLAG-KRAS WT APs grown in minimal medium). Comparing the 2265 high-confidence PPIs determined in this work to 811 previously reported KRAS PPIs (reviewed in [29]) shows an overlap of 245 proteins (hence, 30.2 % of the literature PPI are among the 2265 identified here) and 2020 proteins were not previously reported (Fig. 5.1B). Similar overlaps are obtained when focusing only on the classical effector proteins (Fig. 5.1C).

To gain insights into the whole dataset, a principal component (PC) analysis and a uniform manifold approximation and projection (UMAP) [137] were performed with all high-confidence interactors identified in each AP-MS experiment (Fig. 5.1D; Fig. 5.S5). Both techniques enable a dimensionality reduction of the data and a data visualization. The PC analysis, which is commonly used, tries to preserve the global structure of the data (Fig. 5.1D), while the UMAP tries to preserve the data's local structure (Fig. 5.S5). The unstimulated and KRAS WT samples cluster separately from the other groups suggesting that KRAS WT and unstimulated conditions are a good control group (KRAS mutants and stimulated conditions; middle left area in Fig. 5.1D). Interestingly, KRAS interactor proteins detected in the different mutant datasets cluster together,

compared to the different culture context datasets, where the data are more discriminated. For example, IL-6 and PGE2 context cluster together at the top right corner, whereas DMOG context at the bottom right corner (Fig. 5.1D). Taken together, these results suggest that the proteins detected in complex with KRAS seem to be more conditions-dependent rather than mutation-dependent.

Binding landscape of effectors in complex with KRAS WT and mutants

Effector proteins bind Ras in the GTP-bound state and they are likely forming, among other proteins, the first layer of interacting proteins. Hence, we first characterized the KRAS-effector layer in more detail. As previously mentioned, no effectors are found in complex with KRAS WT in minimal medium, which is expected as KRAS will be mainly in the GDP-bound state that does not enable high affinity binding. Eleven out of 56 classical Ras effector proteins were identified in at least one of the 44 conditions (Fig. 5.2AB). All effectors identified in complex with KRAS belong to either group 1 (AFDN, ARAF, RAF1, BRAF, RGL2) or group 2 effectors (RIN1, PIK3CA, GRB7, RIN2, PIK3C2A, ARAP1). They are generally highly expressed in colon tissue and Caco-2 cells (with medium or high affinities for Ras-GTP) or are moderately expressed but have high affinities in complex with Ras-GTP (Fig. 5.2B). Concerning the 45 effectors not found in any of the KRAS AP-MS samples, 15 belong to group 3 effectors (likely no “true” Ras effectors) and 26 belong to group 2 (of which 7 have low mRNA levels in Caco-2; <3.3 nTPM). Four effectors belong to group 1 effectors, of which RALGDS and RASSF5 are part of the KRAS literature interactome literature and are highly/moderately expressed in Caco-2 cells. SNX27 and RASSF7 are also highly/moderately expressed in Caco-2 cells, but their affinities in complex with Ras are lower. Altogether, we provide a near-to-complete binding landscape of effectors in complex with KRAS under the conditions tested.

A comparison with the effectors identified in previous KRAS interactomes (reviewed in [29]) (Fig. 5.2CD) shows that the AP-MS experiments in this study specifically increase the percentage of group 2 effectors, but little increase in group 1 effector coverage and no increase in group 3 effectors (Fig. 5.2E).

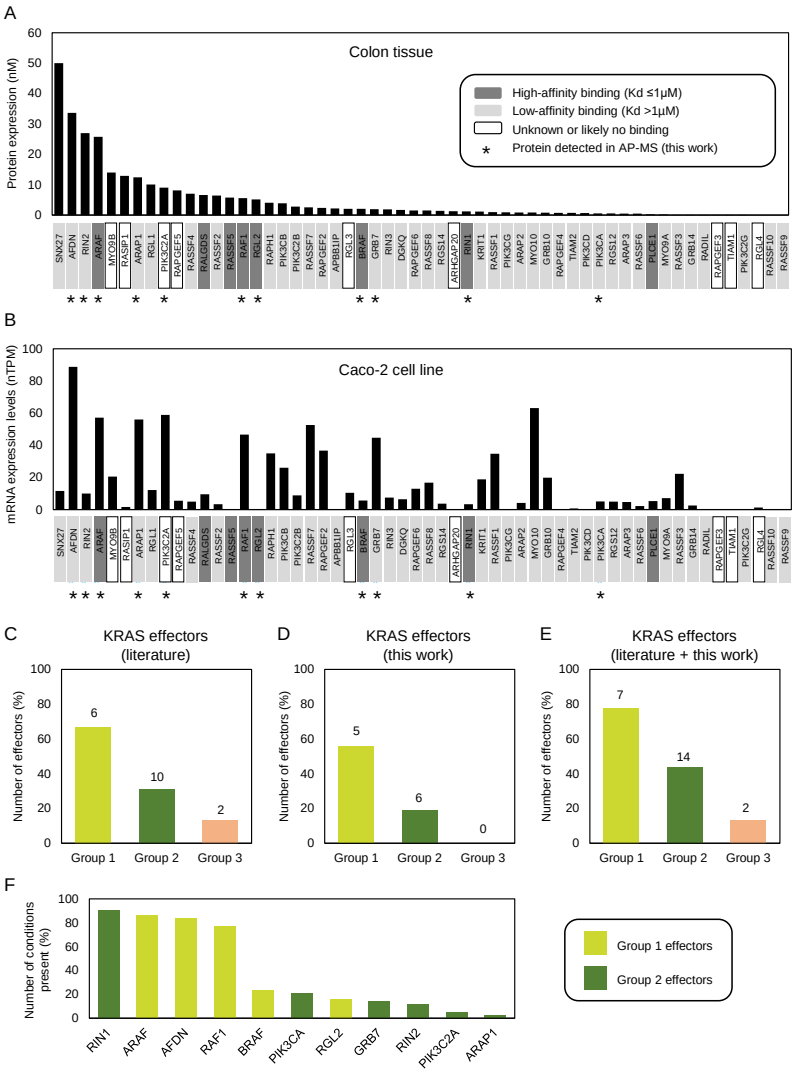

Figure 5.2 Ras effector abundances, binding affinities, and detection in AP-MS experiments. Ras effector protein abundances in human colon tissue based on [60] and [30] (A) and mRNA abundances in Caco-2 cells based on the Human Protein Atlas database (B). The colors of the effectors' names at the bottom of each histogram correspond to Ras-effector binding affinities. The black star indicates effectors that were detected in at least one of the AP-MS experiments conducted in this work. TPM: transcripts per million; nTPM: normalised transcript expression values per sample; K_d: dissociation constant. (C) - (E) Effectors identified in at least one condition in the literature KRAS interactome described in [29] (panel C), in this work (panel D), and in the combined datasets (panel E). Group 1, group 2 and group 3 effectors are normalized based on the total number of effectors in the respective group. (F) Number of conditions where an effectors is present.

Further, the number of conditions in which an effector is identified in complex with KRAS tends to be lower for group 2 effectors (Fig. 5.2F). To visualize in which genetic and culture contexts the 11 effectors were detected, two heatmap images were generated (Fig. 5.3). The two heatmaps show a similar pattern in terms of effector detection and abundances in each of the groups of the AP KRAS mutants (i.e., G12D, G12C and G12V) in unstimulated and stimulated conditions. More specifically, the effectors AFDN, ARAF, RAF1 and RIN1 are detected in all the KRAS mutant AP-MS experiments when unstimulated and stimulated (Fig. 5.3A). These effectors are also detected in the WT AP-MS experiments - albeit not in all the culture contexts. They are classified in the effector group 1 except for RIN1, which is part of the effector group 2. Moreover, they appear to be more abundant when detected in particular conditions such as DMOG and TNF-α, compared to other conditions such as IL-6 or PGE2 (Fig. 5.3B). As these effectors are detected consistently in the presence of KRAS with or without stimulations, we propose that these effectors are KRAS-specific rather than condition(stimulation)-specific. Other effectors are only detected in specific stimulated conditions and are mainly detected in the predicted effector group 2. Moreover, GRB7 is only detected in the presence of DMOG culture context. Another effector, PIK3CA, is only detected when stimulated with TNF-α in the presence of either the G12D or G12V KRAS mutations. The effector PIK3C2A is only detected significantly in the presence of the G12D KRAS mutation with DMOG. These effectors grouped in effector group 2 can be classified as conditions-specific. To mention, the effector BRAF, which was computationally predicted to be always in complex with KRAS, is found in complex to KRAS only in DMOG and TNF-α conditions.

Altogether, this supports our initial hypothesis that group 2 effectors tend to be found in complex with Ras only in specific conditions that promote PM recruitment via RBD-independent mechanisms (Fig. 5.S6). Further, it validates our computational-based classification into group 1, group 2, and group 3 effectors [15] and supports its applicability beyond the 29 human tissues as the basis of the prediction model.

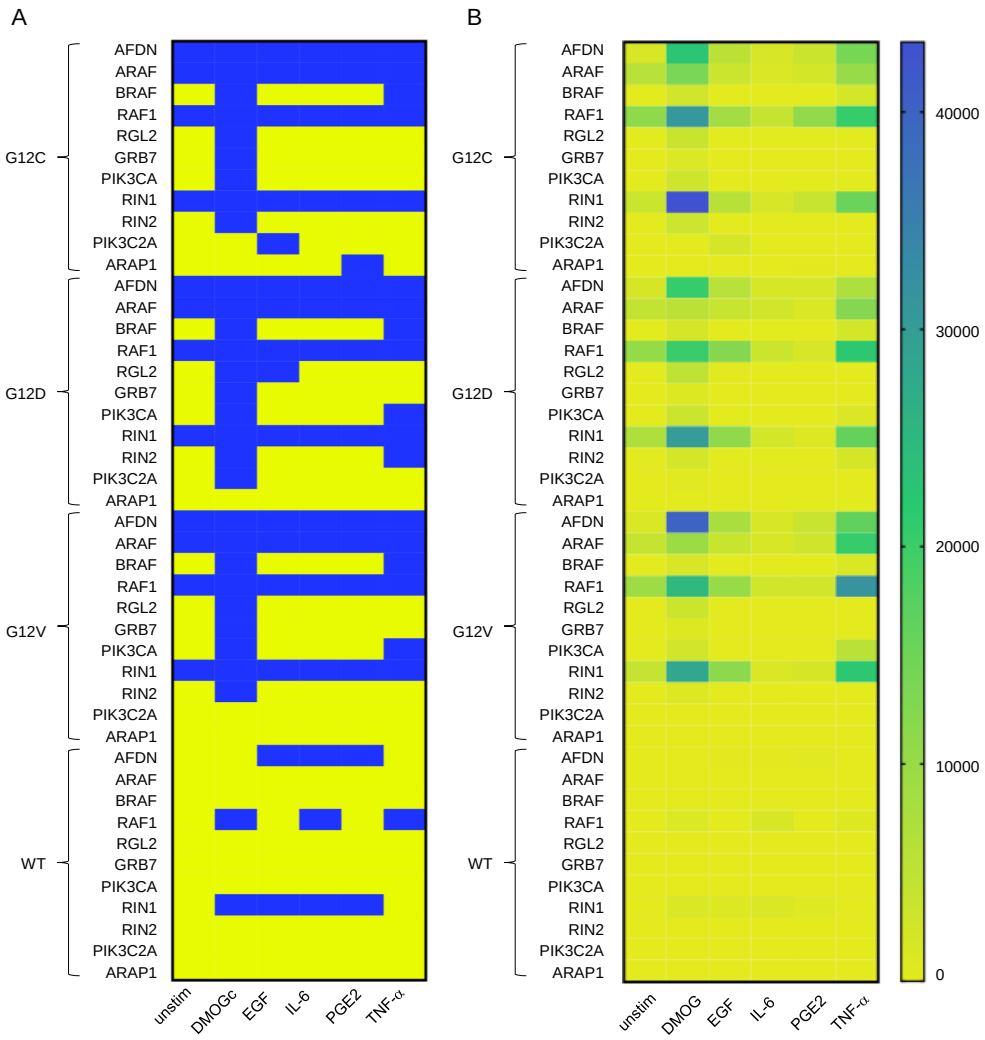

Figure 5.3 Summary of effector presence in AP-MS experiments in Caco-2 cells. The rows display the effectors in complex with KRAS grouped by mutational status (WT, G12D, G12V and G12C) and the columns represent the conditions (unstimulated or stimulated with either DMOG, EGF, IL-6, TNF-a and PGE2). To generate the two heatmaps, the LFQ intensities were analyzed, the different concentrations for each stimulation merged and translate for the heatmap (A) in terms of presence or absence of an effector (color code: detected = blue and not detected = yellow) and for second heatmap (B) the LFQ intensities were directly plotted into the heatmap (code: from low abundance = yellow to high abundance = blue). The heatmaps were created using GraphPad Prism 9.

Investigation of functional differences in the KRAS interactome

To investigate functional differences in the interactome of the different genetic and culture contexts, two approaches were chosen. First, a differential interaction analysis was performed on the identified proteins followed by a gene set enrichment analysis (GSEA) against the Gene Ontology (GO) Biological Processes gene ontology (Fig. 5.4A). Secondly, using the ontology, we collapsed each sample onto all ontologies listed under “biological process” by summing up their LFQ intensities. This process preserved the variation in the data (Fig. 5.S7). Then, multiple ANOVAs were used to identify differences between the samples in terms of their summed up LFQ intensities. Out of the 16.000 GO terms tested, we find significant changes for 2135 (Fig. 5.4B; Fig. 5.S8A). Afterwards, semantic analysis was used to organize the significantly changed ontology terms into clusters. The full distance heatmap from the semantic analysis together with the clusters and some of the data projected onto it is shown in Fig. 5.4. The biggest semantic clusters are linked to metabolic and biosynthetic processes (cluster 1), signaling and immune response (cluster 2), vesicle-mediated transmembrane and ion transport (cluster 3), differentiation, development and morphogenesis (cluster 4), and actin and cytoskeleton organization (cluster 5) (Fig. 5.4B, Fig. 5.S8D-H). Smaller clusters are linked to thermogenesis, ion homeostasis, cell cycle, leukocyte activation, regulation of GTPases, proliferation, and apoptotic cell death (Fig. 5.4B). Altogether, the overall functional differences in the KRAS interactome are consistent with known cellular functions mediated by KRAS [30, 57].

The above two-fold analysis shows different aspects of functional differences between the different genetic and culture conditions. To make the data better approachable, we developed an interactive R Shiny app for exploring the functional terms that are different between the samples (Fig. 5.S9). Users can explore the analysis through the semantic distance heatmap and the semantic clusters, or search and filter for functional terms of their interest, visualize which proteins are part of this particular GO term and show their abundance in the different samples. It also directly displays the samples that are statistically significant to each other. We propose this app as a resource to filter for

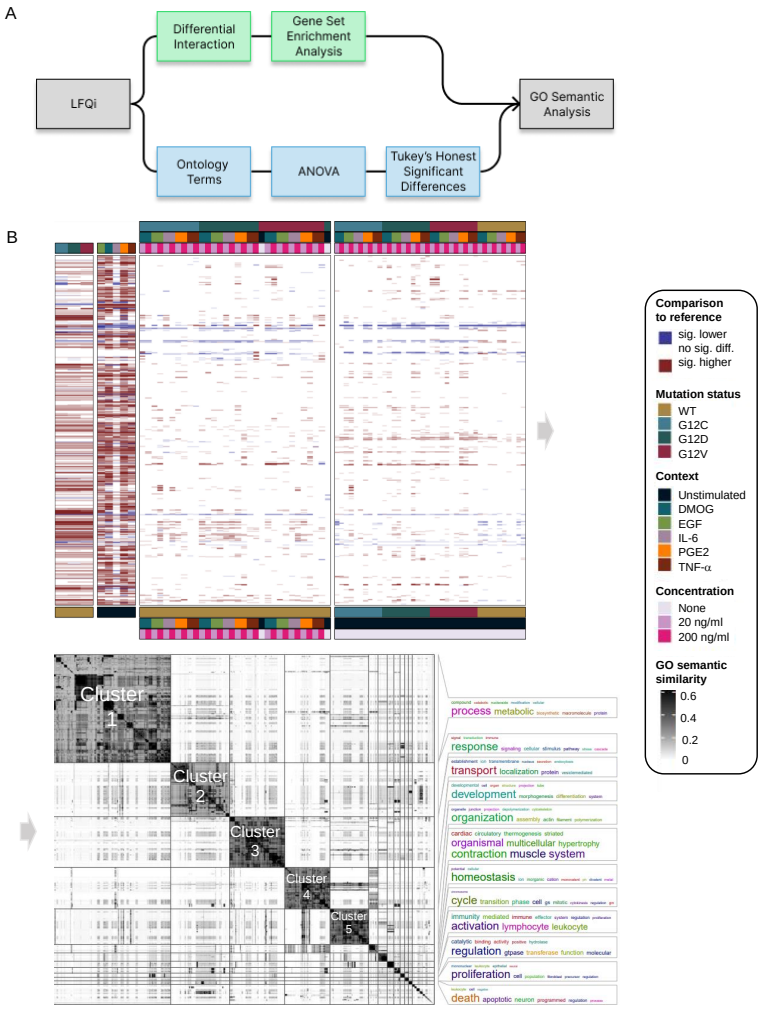

Figure 5.4 Gene set enrichment analysis (GSEA) of AP-MS analysis in different genetic and culture contexts. (A) Overview of the functional analysis pipeline. The first approach is a standard differential analysis pipeline using limma, followed by a gene set enrichment analysis against the "Biological Process" GO ontology. The second approach is to sum up LFQ intensities for each of the "Biological Process" GO ontology terms. Then, for each term, an ANOVA was performed to identify whether there was a significant influence of mutation status, condition and concentration and their interaction terms. Both analysis were then analyzed using semantic analysis of ontology terms. (B) Distance matrix with the pairwise semantic distances between the GO terms is shown in the center of the plot. On the right, the bigger clusters are annotated by word clouds. On the left, the data from the ANOVA and the GSEA analysis pipeline is shown, for the ANOVA whether a main effect is statistically significant and for the GSEA, whether a significant enrichment was found in the relevant pairwise comparison. All data shown is in comparison to the WT and unstimulated for the analysis of mutation status and condition, respectively.

interesting functional influence of certain KRAS mutations or certain growth conditions from our data set. From the results of this functional investigation, we selected some GO terms of interest to us which we went on to validate in the wet lab. Those were GO terms related to proliferation ("Epithelial cell proliferation" and "Positive regulation of cell population proliferation"), glucose metabolism ("Glycolytic process" and "Regulation of glucose metabolic process"), and ATP metabolism ("ATP metabolic process", and "Regulation of ATP metabolic process") shown for the genetic contexts G12C and G12D and the culture contexts DMOG and IL-6 in Fig. 5.5A). As we are studying the effect of oncogenic KRAS-mutation in an adenocarcinoma cell line under CRC microenvironment mimicking culture contexts, we were particularly interested in biological processes commonly observed in cancer development. Among the hallmarks of cancer [164], sustained cell proliferation and the Warburg effect, described by a switch in cell energetic metabolism from oxidative phosphorylation to aerobic glycolysis, are two biological and metabolic processes that are feasible to test experimentally. In the GO functional analysis, we observed that PGE2 and DMOG had the greatest number of differentially expressed GO compared to unstimulated while IL-6 had the lowest. In addition, we observed that the differential expression profile of the three KRAS mutants were similar, and KRAS G12C and G12D differ the most for GO terms related to biological and metabolic processes.

Analysis of cell phenotypes in selected genetic and culture contexts

To investigate changes in cell proliferation and glycolytic metabolism for three genetic contexts (KRAS WT, G12D and G12C) in three culture contexts (unstimulated, DMOG and IL-6), we measured cell count, cell viability, glucose uptake and lactate release over a 72 h period in Caco-2 cells (see methods). We chose to determine Caco-2 metabolism and proliferation until 72 h post-transfection under the assumption that the changes observed in the interactome seen 24 h post-transfection would affect cell phenotype on a long-term basis and that phenotypic changes can take longer to be accurately measured. To ensure that exogeneous KRAS was expressed until 72 h post-transfection we performed

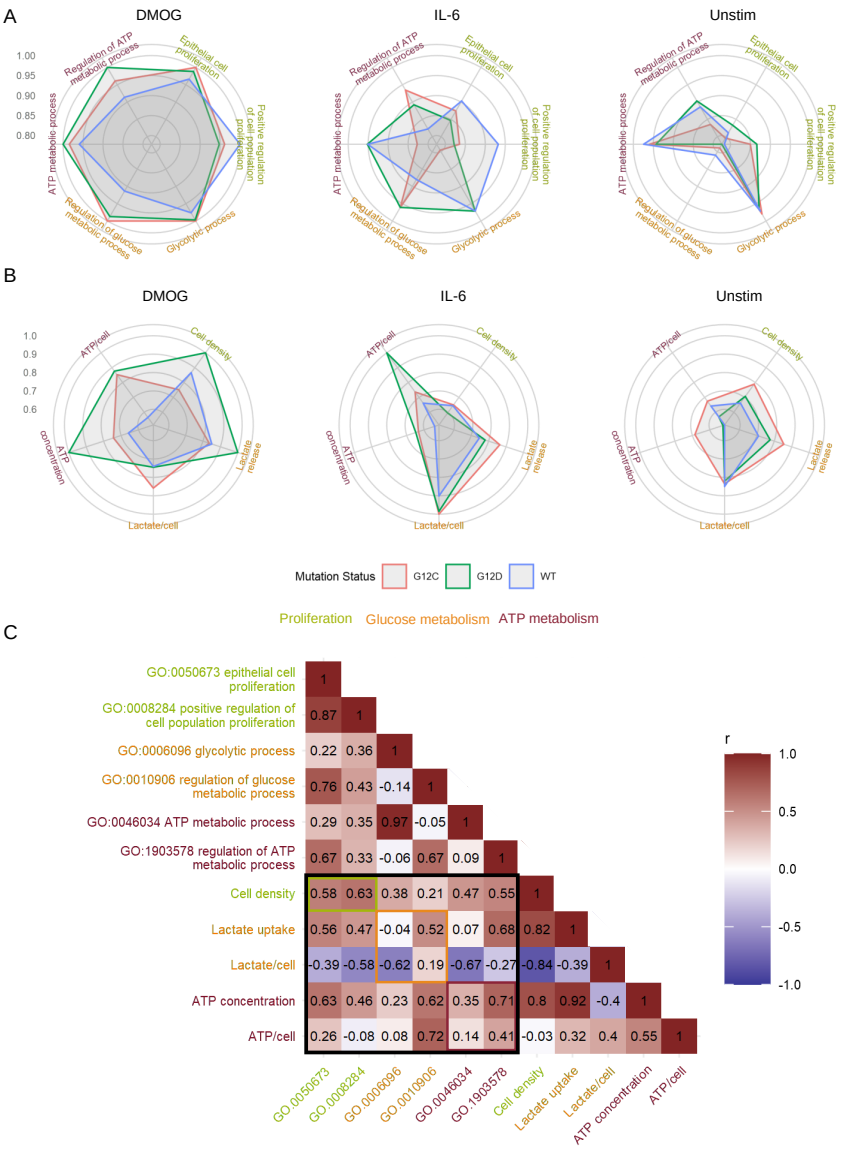

Figure 5.5 Comparison of phenotypic parameters with the sum of LFQ intensity of related GO terms in the AP-MS. (A) Radar plot of phenotypic parameters related to cell proliferation, glucose metabolism and ATP metabolism of Caco-2 cells under different genetic and culture contexts. (B) Radar plot of the GO terms related to cell proliferation, glucose metabolism and ATP metabolism of Caco-2 cells under different “culture” and “genetic” contexts. Value in radar plots are the LS mean values normalized by the maximum value of each parameter to obtain values between 0 and 1. (C) Pearson correlation matrix of phenotypic parameters versus sum of LFQ intensity of specific GO terms from the AP-MS. Phenotypic parameters and GO terms related to the same biological process are displayed with the same colors.

a Western-Blot of FLAG-KRAS before transfection and from 24 h to 72 h post-transfection. We observed an overall significant increase of FLAG expression 24 h post-transfection with a peak at 48 h, and then a decrease of FLAG expression at 72 h to reach the level of 24 h time point (Fig. 5.S10). Those results suggest that for all genetic contexts, exogenous KRAS was effectively expressed during a 72 h period. However, we also noticed that exogenous KRAS expression was significantly different between the three genetic contexts.

We observed a significant increase in cell proliferation under DMOG context compared to the control (unstimulated), although IL-6 led to a significant decrease in cell proliferation (Fig. 5.S11A). The increase in cell proliferation in DMOG culture context was confirmed by cell viability results (ATP concentration, Fig. 5.S11G). Furthermore, for those two phenotypic parameters, under DMOG stimulation, KRAS G12D had a significantly greater cell proliferation than KRAS WT and G12C. In terms of glycolytic metabolism, it was overall better captured by lactate release than glucose uptake and the highest lactate release was observed in DMOG context. A greater lactate release was also observed for both KRAS mutants compared to WT (Fig. 5.S11E). When normalized per cell, both glucose uptake and lactate release were significantly greater in the IL-6 condition than in the control and DMOG, suggesting a higher glucose utilization at the single cell level in IL-6 context. Finally, the ATP pool per cell, which can be used as a proxy for changes in metabolic activity, was higher for both stimulations compared to the control and for both oncogenic KRAS compared to the WT. Altogether, the experimental results suggest that under genetic and culture contexts that mimic colon and CRC, Caco-2 cells produce more energy than KRAS WT cells in control conditions. It also suggests that changes on the level of the interactome, particularly those seen with different culture (microenvironmental) contexts, likely manifest as phenotypical changes. It supports our initial hypothesis that culture contexts are modifiers of both KRAS-mediated networks and downstream phenotypes.

Next, we compared the measured phenotypical parameters to the sum of LFQ intensities of proteins associated with GO terms linked to proliferation, glucose metabolism, or ATP metabolism with either term directly related to the final cell phenotype or its regulation (Fig. 5.5). Overall, the culture context

effects are similar between the AP-MS (GO term) and the phenotypical experiments. Indeed, we observed that for all GO terms excepts “Glycolytic process” the sum of LFQ intensities were significantly higher in DMOG compared to unstimulated and IL-6 (Table 5.S1). Those results are similar to the cell density, lactate release, and ATP concentration experimental results. In addition, for “Regulation of glucose metabolic process” the sum of LFQ intensity was higher in the IL-6 context compared to unstimulated, which could be linked to the higher lactate release per cell observed in the IL-6 condition. The only notable difference is observed for the “Glycolytic process” where the sum of LFQ intensities was significantly lower in the IL-6 culture context, which is contrary to what was observed for lactate release per cell. Results of the effect of the genetic context are also to some extent similar, with no significant effect on proliferation related GO terms such as cell density. In addition, a significantly higher sum of LFQ intensities was observed in the “Regulation of glucose metabolic process” in the two KRAS mutant proteins compared to WT, with results similar to the lactate release. However, no genetic context effect was observed for ATP metabolism-related GO terms, although mutant KRAS had a higher overall ATP concentration than WT. Especially, the sum of LFQ intensities of “Glycolytic process” was significantly lower in G12C than in the two other genetic contexts mostly due to a substantial decrease under IL-6, which is opposite to what was found for overall lactate release. Those results suggest first that the strongest changes in phenotype, such as those observed for DMOG stimulation, are more reliably captured by the AP-MS than the milder changes such as those observed for genetic context or IL-6 stimulation. Second, the changes observed for GO terms related to the regulation of a process seem to be better at capturing the results of phenotypical experiment than GO terms related to the biological process itself. This latter point is also suggested by the better Pearson’s correlation coefficient observed for phenotypical parameters with the associated GO terms related to the regulation of a biological process than those directly related to the biological process (Fig. 5.5C). This may be because proteins that have a direct effect on phenotype are too many layers downstream of RAS signaling network to be precisely captured by AP-MS experiments. Finally, for metabolic parameters, the functional analysis results

are more correlated with overall lactate release and ATP concentration than with the values normalized per cell. This implies that the results of the functional analysis based on AP-MS data are more suitable at capturing changes in metabolism at the level of the whole cell culture than at the level of a single cell.

Information flow analysis predicts the contribution of effectors to functional processes

After demonstrating that the differences in the summed LFQ intensities for a functional process can be associated with functional differences in the behavior of the Caco-2 cells, we aimed to further explore how the changes in the interactome led to these functional differences. In particular, we were interested in understanding which effectors and proteins downstream of KRAS were involved for a specific process. To this end, we used random walks over a filtered version of the STRING network, in which we biased the random decision for each step depending on whether a potential next protein was part of the AP-MS dataset for a specific sample or not (see methods). This left us with a collection of paths and their probability to be traversed, based on the network architecture and the proteins found in the AP-MS sample.

Applying this method to the functional terms and samples we were interested in, we observed drastic changes in the network architecture depending on the sample (Fig. 5.6). Particularly, for some GO terms there is a different predicted engagement of effectors. Examples are the two GO terms related to “Epithelial Proliferation” for KRAS oncogenic mutations in culture contexts IL-6, DMOG and unstimulated (Fig. 5.S12). For “Epithelial cell proliferation”, in IL-6 context high path counts are found for the effectors AFDN, ARAF, and RAF1 (Fig. 5.6A). In DMOG context, additional high counts are found for BRAF, GRB7, and PIK3CA (Fig. 5.6A). Likewise, for “Positive regulation of cell population proliferation” contributions of AFDN and PIK3CA dominate in DMOG context (Fig. 5.6B). For GO terms related to glucose metabolism, AFDN, GRB7 and PIK3CA dominate the path counts for “Glycolytic process (GO:0006006)” in DMOG condition and for KRAS G12C in IL-6 context, but has a low count

for “Regulation of glucose metabolic process (GO:0010906)”, where ARAF, BRAF and RAF1 dominate in most culture conditions (Fig. 5.6CD; Fig. 5.S13). With respect to GO terms related to ATP metabolism, the culture contexts IL-6 and DMOG show profound path count differences for the effector RAF1, which contributes more in DMOG than in IL-6 context (for KRAS G12D genetic context), but slightly higher path count in IL-6 context for KRAS G12C (Fig. 5.6EF; Fig. 5.S14). The path count is generally low for BRAF and ARAF. AFDN has high/the highest path counts in almost all genetic and culture contexts. Also noteworthy, RIN1 is present in all contexts but not necessarily much involved. Altogether, our biased random walk analyses predicts the contribution of individual effectors to GO terms that link to experimental phenotypes.

5.2.4 Discussion

This study set out to explore KRAS as a key cellular signaling hub in specific relevant (patho)physiological contexts. The Caco-2 cell line has been used as a relevant model system that can be grown in various growth media (culture contexts) and enable exogenous expression of KRAS WT and oncogenic mutants (genetic contexts). Indeed, Caco-2 cells are human intestinal epithelial cells that closely mimic the colon intestinal epithelium in the early stage of CRC. By identifying different levels of network organization (e.g. sub-complexes and number of paths traversing a network), we aimed to detail the downstream pathway of KRAS further and investigate the functional outputs. To address this challenge experimentally, even though all methodology has their limitations, AP-MS excels for profiling interactomes in humans [161] due to its sensitivity and its ability to detect interactions within complexes in appropriate contexts [165]. We successfully pulled down complexes by using the exogenous expression of tagged bait proteins for different KRAS mutations on Caco-2 cells.

Effectors bind to KRAS in a mutually exclusive fashion and can potentially compete for binding [110]. Our earlier computational predictions suggested that there is a considerable impact of the culture contexts on the recruitment of specific effectors to the PM [15]. Here, we identified a total of 11 effectors in at least one of the AP-MS experiments, of which seven are only found in complex

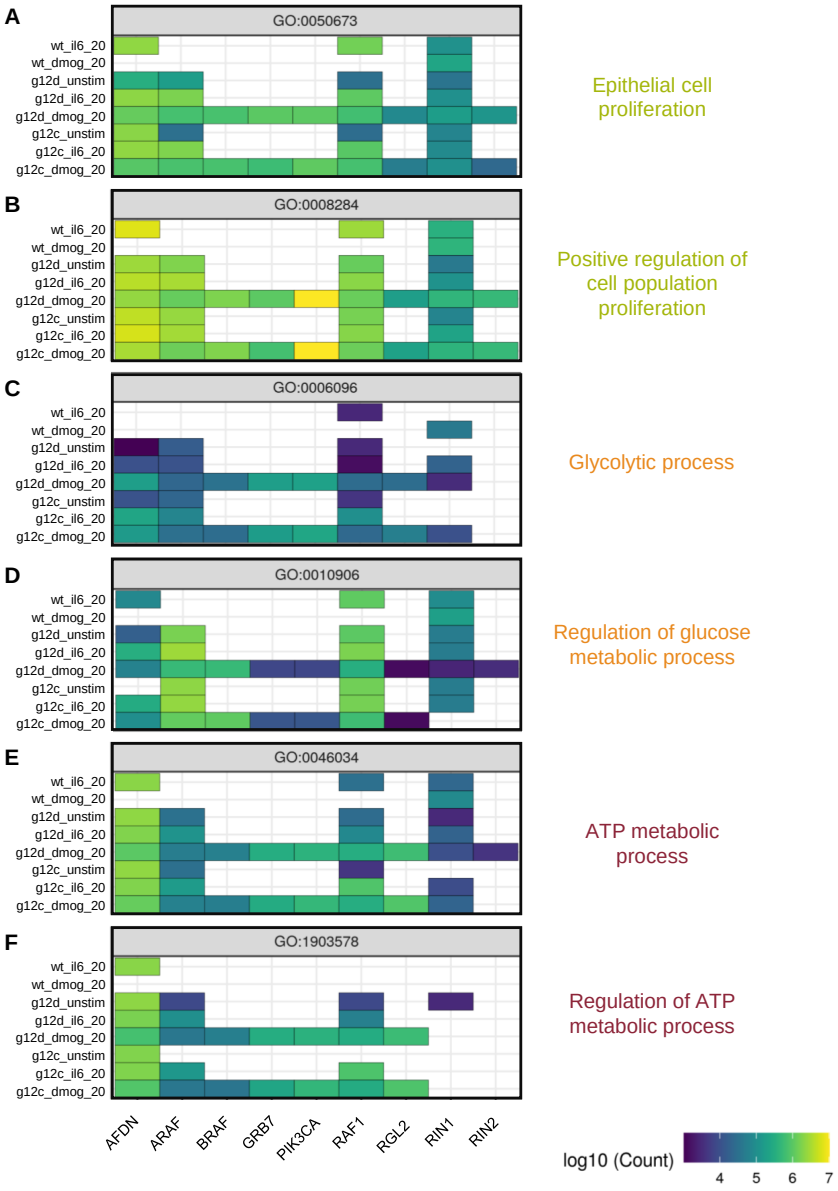

Figure 5.6 Pathway analysis by biased random walks. (A) Heatmaps of effector traversal in the different genetic and culture for selected GO terms associated with cell phenotypes. Epithelial cell proliferation (GO:0050673). (B) Positive regulation of cell population proliferation (GO:0008284). (C) Glycolytic process (GO:0006096). (D) Regulation of glucose metabolic process (GO:0010906). (E) ATP metabolic process (GO:0046034). (F) regulation of ATP metabolic process (GO:1903578). The colour scale indicates how often an effector is traversed.

with KRAS in some genetic and culture contexts. For example, effectors such as PI3KCA, RIN2, GRB7 or ARAP1, are detected in the presence of culture conditions such as hypoxia (HIF-stabilization), IL-6, and TNF-α. We predict that in these cases the affinity between the RBD and KRAS is not high enough to allow for sufficient binding and that additional domains present in effectors are required to increase the number of complexes formed between KRAS and effectors at the PM (based on the "piggyback" mechanisms; [111]). Indeed, we show in this work that the total number of effectors and other proteins in the Ras-mediated complex increases with the number of conditions. This context dependent binding can be explained by the fact that cells in their physiological microenvironment are constantly experiencing a variety of stimuli that trigger receptors (e.g. the EGF receptor) located on the plasma membrane, where Ras is located [166]. In the context of cancer, tumor cells are often located in a hypoxic, immunosuppressive and nutrition-deficient microenvironment that causes reprogramming of metabolism and signaling [164]. Indeed, we identified culture context-specific metabolic alterations in glucose and ATP metabolism in the Caco-2 cells. Hence, this work supports the requirement to study the role of the microenvironment when performing an experiment that aims to characterize PPI networks because they have a major role in driving complex formation. It also demonstrates the need to consider multidomain interactions. However, the interpretation of PM recruitment and culture conditions might not be straightforward. In fact, it is difficult to understand and predict what happens on the upstream level of Ras and effectors and why some effectors are identified in the specific conditions tested. This is partly due to signaling pathways that are highly cell type-specific [140, 167]. Together with different databases such as the Human Protein Atlas initiative [168] and the large-scale interactome Bioplex based on AP-MS baits [169, 170], analyzing human PPI and the conditions in which they occur is essential for the development of a context-dependent human interactome. Indeed, a new version of the Bioplex 3.0 interactome has been recently published where, in addition to the HEK 293T cells, a dual comparison with the HCT116 cell line was performed [165].

Based on the assumption that effectors compete for binding to KRAS, our working hypothesis is that individual KRAS-effector-mediated sub-complexes

form in a cell, which ultimately affect downstream signal propagation and cellular phenotypes. Indeed, we show here that differences in KRAS-mediated complexes propagate to downstream changes in phenotypes that roughly align with the predicted functional changes based on GO terms of proteins detected in an AP-MS experiment. This suggests that the PPI network orientation/assembly on the level of KRAS (likely mediated in part by effectors in complex with KRAS) impacts the downstream phenotype. As effectors compete for binding to KRAS we hypothesized that specific Ras-effector sub-complexes exist that each (or in combination) link to specific phenotypes. To explore the contribution of individual effectors to phenotypes, we used biased random walk analysis. Indeed, we find differences in the number of paths between different genetic and culture contexts. The analysis also enabled us to predict which effector pathways are likely linked to cellular phenotypes. Hence, our analysis pipeline that combines AP-MS data with random walks and GO terms provides a novel way to link PPI networks to phenotypes. The pipeline and code is available to the scientific community and can be adapted for specific AP-MS experiments. There are, however, limitations of the random walk analysis. The data structure at the end is a collection of different paths for different targets for different conditions. Some of these are comparable, some are biased. Paths ending into the same target should be comparable across conditions, as long as the underlying network structure does not change. However, there is a bias for shorter paths to be more likely to be found, and targets with shorter shortest-paths have on average higher counts in the found paths. Additionally, the analysis is strongly dependent on the underlying network structure that is used.

With respect to the impact of genetic vs culture contexts on KRAS-mediated network rewiring, our analyses based on PCA and UMAP suggest that the impact of growth condition (culture context) is greater than type of oncogenic mutation (genetic context). We observe a similar trend in the functional analysis as well as in the effector contributions as calculated by random walks, where different oncogenic mutants generally having more similar effector path counts are different culture contexts/ growth conditions. Indeed, the results of this work offer an additional explanation why cancer genes and mutations only manifest in some but not all tissues [171]. Future steps in systems medicine require the

integration of protein abundances with context-specific conditions and localized signaling responses. Indeed, quantitatively predicting the influence of specific conditions on larger networks to get an efficient predictive model would be ideal, especially in the case of oncogenic mutations. In addition, understanding the rewiring in physiological contexts to enhance the understanding of network rewiring in cancer contexts would provide new insights into potential therapeutic targets [172].

5.2.5 Materials and Methods

Culturing of Caco-2 cells

Caco-2 cells (ATCC©HTB-37) were cultured in Dulbecco's Modified Eagle's Medium (DMEM) (GibcoTM, ThermoFisher, 21969-035) supplemented with 2 mM L-glutamine (GibcoTM, ThermoFisher, 25030-024), 10% (v/v) Foetal Bovine Serum (FBS) (GibcoTM, ThermoFisher, A4766801), and 1% Penicillin/streptomycin (GibcoTM, ThermoFisher, 15140122). For long-term storage, frozen stock vials were made on the week of receiving the cell line in RecoveryTM cell culture freezing medium (GibcoTM, ThermoFisher, 12648010) and stored in liquid nitrogen. For each experiment, cells were not exceeding passage 25 and were thawed from the liquid nitrogen stock. To generate growth media that mimic conditions relevant in the colon and CRC ("culture contexts"), the minimal medium (DMEM with 2 mM L-glutamine) was supplemented with either IL6 (interleukin-6) (ThermoFisher), TNF-α (tumor necrosis factor-alpha) (ThermoFisher), PGE2 (prostaglandin E2) (ThermoFisher), EGF (epidermal growth factor) or the HIF-Hydroxylase Inhibitor DMOG (Cayman chemical) at different concentration (20 and 200 ng mL^{-1}).

Plasmids for exogenous expression of FLAG-KRAS WT and mutants

Plasmids were gifted from the previous research laboratory of Christina Kiel in Barcelona (CRG) (from Luis Serrano and Hannah Benisty). All the plasmids harbor the identical backbone pMDS-TetOn3G-kozak-FLAG-GOI (gene

of interest). Plasmids differ only by their GOI, which are wildtype (WT) KRAS, KRASG12D, KRASG12V, or KRASG12C as GOI.

Bacterial transformation with plasmids, plasmid extraction and purification

The bacterial transformation of the plasmids for exogenous expression of FLAG-KRAS (pMDS-TetOn3G-kozak-FLAG-KRAS WT/mutants) was performed using the One-Shot™ Stbl3™ (Invitrogen, C737303) chemically competent bacterial cells to replicate each plasmid following the manufacturer's instructions. Subsequently, 100 µL of the bacteria-plasmids solutions were plated into LB selective agar plates containing the antibiotic spectinomycin (50 µL). Plates were incubated at 37 °C, overnight. The next day, individual bacterial colonies were selected from LB agar plate and grown in 4 mL LB broth with the corresponding antibiotics for 6 to 12 h at 37 °C in a shaking incubator at 250 rpm. After incubation, several aliquots of this original starter culture were used to generate a bacterial glycerol stock for long-term storage at –80 °C (1 mL transformed bacteria in 1 mL 50% glycerol). The remainder of the original starter culture was then used to grow at a large scale the transformed bacteria under selective antibiotics overnight in 500 mL of LB medium at 37 °C in a shaking incubator at 250 rpm. The HiSpeed Plasmid Maxi Kit (Qiagen) was used to generate larger amount of the FLAG-KRAS plasmids. The kit was used following the manufacturer's instructions. Final DNA was eluted in 400 µL of TE buffer and allowed to resuspend overnight to ensure homogeneity. The following day, concentrations and purities were measured on the Implen NanoPhotometer®NP80, and plasmids DNA were stored at–20 °C.

Transfection and expression of FLAG-KRAS in Caco-2 cells

For AP-MS experiments, Caco-2 cells were seeded 24 h before transfection in 10 cm dishes in normal growth medium and grown to 70-80% of confluency. Cells were transfected with 15 µg of pMDS-TetOn3G-kozak-FLAG-GOI plasmids (containing FLAG-KRASWT or FLAG-KRASG12D or FLAG-KRASG12V

or FLAG-KRASG12C as GOI) using Lipofectamine 2000 (Invitrogen, 11668-019) according to the manufacturer's instructions in OPTI-MEM reduced serum medium (GibcoTM, ThermoFisher, 31985-062) for 4 h. Then, the medium was changed and supplemented with culture medium containing the various growth conditions. To note, cells transfected with the KRASWT plasmids were always supplemented with 15 ng mL^{-1} of doxycycline (Sigma-Aldrich). Cells were incubated for 24 h at 37 °C and harvested. FLAG-KRAS mutant plasmid transfections were not supplemented with doxycycline as the promoter is leaky and KRAS mutant proteins were already expressed at WT levels without adding doxycycline.

Caco-2 cell lysis, protein extractions and concentration

Caco-2 cell lysates were obtained after trypsinization, and cell pellets were recovered and washed twice with PBS 1X. The cells pellets were then resuspended in the appropriate volume (e.g., 300 µL for the AP-MS experiments) of lysis buffer (50 mM TRIS HCL pH 7.5, 1 mM EDTA, 1 mM EGTA, 150 mM NaCl, 2 mM MgCl2, 1 mM DTT, and 1% IGEPAL/NP-40 supplemented with PhosSTOP (Roche) and cOmplete, Mini protease inhibitor cocktail (Roche)). Cells were lysed for 30 min on a rotator at 4 °C, centrifuged at 14 000 rpm for 30 min at 4 °C, and the supernatants were collected in a new tube. Protein concentrations were measured using the PierceTM 660 nm Protein Assay (ThermoFisher, 22660) per manufacturer's guidelines. Samples were incubated for 5 min before absorbance was read at 660 nm on a SpectraMax M3 plate reader. Net absorbances were plotted against BSA protein concentration for standard curve generation (ThermoFisher, 23208). For each sample, the concentration was obtained by comparing net absorbance values against the generated standard curve. A new standard curve was generated for each assay. Kept on ice, cell lysates were then directly used for affinity purification.

Western blotting

Prior to loading the samples into the gel, a normalization of the concentration for each sample is done, with a concentration aiming to be 1 µg µL^{-1}. Proteins were then denatured by incubating samples at 95 °C for 5 min in 4× Laemmli buffer and dithiothreitol (DTT) before loading onto 4 to 12% NuPAGE gradient precast gels (ThermoFisher). Gels were run for 10 min at 110 V, followed by 45 min at 150 V, with gels submerged in NuPAGE MES running buffer (ThermoFisher). After electrophoresis, proteins are dry-transferred using the iBLOT2 device (ThermoFisher) for 7 min into a nitrocellulose membrane. The membranes were checked by Ponceau S staining to ensure protein transfer. Then, the membranes were washed in 1× Tris buffer saline-tween20 (TBS-T) before blocking solution for 1 h in 5% milk at room temperature in a shaking device. Depending on the antibody, the membranes were incubated either overnight at 4 °C or at room temperature for 4 h, with the primary antibody diluted in 0.05% milk in TBS-T. The membranes were then washed three times in TBS-T, 10 min each and incubated horse radish peroxidase (HRP) conjugated secondary antibody for 1 h diluted in milk TBS-T on a shaking device. Protein bands were developed using high sensitivity ECL reagent (ThermoFisher) with the West Pico western blotting substrate per the manufacturer's instructions and visualized using the G-Box image developer (SYNGENE). Densitometry analysis was performed using ImageJ, with target protein bands normalized to loading control (β-actin or GAPDH). The following antibodies were used for Western blotting: beta-Actin (Cell Signaling, #4970, rabbit / monoclonal, 1/3000 dilution), GAPDH (Abcam, ab2118, rabbit / monoclonal, 1/1000 dilution), panRAS (Abcam, ab52939, rabbit / monoclonal, 1/5000 dilution), KRAS (CPTC-KRAS4B-2, DSHB, mouse / monoclonal, 0.5 µg mL^{-1} working concentration), and secondary Anti-mouse HRP (Abcam, ab97023, goat / monoclonal, 1/3000).

Affinity purification (AP)

Caco-2 cell lysates expressing FLAG-KRAS proteins were immunoprecipitated from 800 µg of cell lysate using anti-FLAG-M2 magnetic beads (Sigma, M8823) by using the KingFisher DuoPrime purification system (ThermoFisher). Beads were washed in TBS (according to the manufacturer's instructions) for 5 min, twice, at low speed. Then beads were collected by the KingFisher magnet and discarded into the samples wells and mixed at a slow speed for 1 h. Beads-antibody samples were collected and went through different wash salted solutions (Wash 1 and 2: RIPA buffer with 150 mM NaCl; Wash 3: RIPA buffer with 500 mM NaCl), mixed at low speed for 30 s. Beads-antibody-sample were eluted in 50 µL of glycine (0.1 M, pH 3.0) for 5 min. Immediately after, samples were neutralized with 20 µL of TRIS BASE (1 M, pH 8.0).

Sample preparation after AP for mass spectrometry (MS)

For protein cleanup the paramagnetic bead–based SP3 (solid-phase-enhanced sample-preparation) workflow was used [173]. For each AP experiment sample protein concentrations were determined using the PierceTM BCA protein assay (ThermoFisher) following the manufacturer's instructions and 50 µg of proteins were adjusted in 20 µL of buffer/MS grade water. Samples were homogenized and denatured in urea (final concentration, 4 M), ammonium bicarbonate (100 mM), and calcium chloride (100 mM), then reduced in DTT (final concentration, 1 mM) for 15 min at room temperature and alkalinized in iodoacetamide (IAA) (3 mM) in the dark at room temperature for 15 min. The tryptic digestion protocol was performed using the KingFisher DuoPrime purification system (ThermoFisher) in a series of steps. First, magnetic hydrophobic and hydrophilic beads were washed several times in MS grade water and added to the deepwells plate in the KingFisher along with the samples and the same volume as the sample of 100% ethanol. Next, the solutions were mixed at low speed for 10 min, after which the beads coupled to the proteins were collected with the magnetic arm of the KingFisher and transferred to be washed in 3 different deepwells containing each 80% of ethanol. The washed beads-

proteins were then released into the trypsin (Promega, V5111)-containing deepwells at a 50:1 (w/w) protein to protease ratio and mixed at low speed for 8 h of digestions into peptide fragments at 37 °C in the KingFisher. Peptide samples were transferred into low protein binding tubes, 1% of trifluoroacetic acid (TFA) was added to acidify the samples ready to be desalted, cleaned, and concentrated on C18Tips (ThermoFisher, 87784) [174] according to the manufacturer's instructions. Purified peptides were dried and resuspended in low protein binding tubes before mass spectrometry analysis in 30 µL of 0.15% TFA acid and 1% acetic acid in mass spectrometry grade water.

Mass spectrometry

The peptides were analyzed using a MS shotgun proteomics technique. This technique allows a sensitive bottom-up approach that consists of separating peptides resulting from protein digestion by liquid high-performance liquid chromatography (HPLC) followed by tandem mass spectrometry (MS/MS). Samples were run on a Bruker timsTof Pro mass spectrometer connected to an Evosep One liquid chromatography system. Tryptic peptides were resuspended in 0.1% formic acid and each sample was loaded onto an Evosep tip. The Evosep tips were placed in position on the Evosep One, in a 96-tip box. The autosampler is configured to pick up each tip, elute and separate the peptides using a set chromatography method [175]. The chromatography buffers used were buffer B (99.9% acetonitrile, 0.1% formic acid) and buffer A (99.9% water, 0.1% formic acid). All solvents are LCMS grade.

The mass spectrometer was operated in positive ion mode with a capillary voltage of 1500 V, dry gas flow of 3 L min^{-1} and a dry temperature of 180 °C. All data was acquired with the instrument operating in trapped ion mobility spectrometry (TIMS) mode. Trapped ions were selected for ms/ms using parallel accumulation serial fragmentation (PASEF). A scan range of (100-1700 m/z) was performed at a rate of 5 PASEF MS/MS frames to 1 MS scan with a cycle time of 1.03 s [176].

The data analysis was done using MaxQuant software [67]. The raw data was searched against the Homo sapiens subset of the Uniprot Swissprot database

(reviewed) with the search engine Maxquant (release 2.0.3.0). Specific parameters for trapped ion mobility spectrometry data-dependent acquisition (TIMS DDA) were used: Fixed Mod: carbamidomethylation; Variable Mods: methionine, oxidation; Trypsin/P digest enzyme (maximum 2 missed cleavages); Precursor mass tolerances 10 ppm; Peptide FDR 1%; Protein FDR 1%. The normalized protein intensity of each identified protein was used for label-free quantitation (LFQ) using the MaxLFQ algorithm [68].

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [177] partner repository with the dataset identifier PXD035399.

AP-MS data filtering and ID mapping

The data were first filtered based on the label-free quantification intensities (LFQi) using the following five steps: (i) removal of proteins that were labeled as “only identified by site”, “potential contaminant”, and “reverse”; (ii) removal of all observations with LFQi equals to 0; (iii) removal of outlier samples (based on low overall LFQi; see Fig. S3); (iv) removal of proteins which are not present in at least 60% of the samples of a group for each group (a group is defined as the collection of three biological with 2 technical replicates for one condition, which results in a group size of maximum 6); (v) filtering against the negative control sample, which is only the beads used for the AP-MS sample preparations, by only considering proteins for further analysis that are significantly higher found in the samples compared to the negative control. In MS analysis based proteomic data, there are typically two types of missing values, the missing not at random (MNAR) and the missing at random (MAR) [178]. A mixed imputation strategy was chosen, with kNN imputation as the strategy for MAR values [179, 180, 181]. Other missing values were considered MNAR values and imputed at value 0. After the imputation, differential interaction analysis was performed for each group against the beads control. P values were adjusted using FDR correction as described by Benjamini-Hochberg [182]. Afterwards, all proteins were extracted for each group which were significantly enriched in the sample (cutoffs: p-value adjusted: <0.01, Log Fold Change: >1). The

data were transformed to have consistent protein and gene names annotations following the data filtering. The data are received from the MaxQuant software in UniProt IDs and mapped to HGNC gene names using the HGNC database (retrieved 12/2021) However, one UniProt ID can correspond to multiple HGNC gene names. In this case, manual selection of the gene names of interest was performed. Finally, the HGNC names were mapped to genes IDs of the SysGO database [183]. A couple of proteins could not be found in the SysGO database, and one protein was renamed (i.e., HGNC name: PHB1, which was renamed PHD for SysGO). Then, the technical replicates were merged using the median. In summary, we obtain a dataset with raw LFQi (Data S1) or log2 transform (Data S2) data with biological triplicates. Data preparation was performed in R ([69], `http://www.r-project.org/index.html`) using the following packages: `dplyr` [83, 184], `tidyr` [185], `stringr` [186], `tidyxl` [187], `purrr` [188], `DEP` [70] and `limma` [71, 72]. The script file for the data preparation as well as the data pre and post preparation are available on Zenodo (`https://zenodo.org/record/6896565`) (DOI: 10.5281/zenodo.6896565).

Functional analysis of the interactome

Functional analysis of the interactome was performed in two different ways. The first approach consists of a differential interaction analysis based on the filtered LFQ intensities. Imputation was performed by a mixed imputation strategy, using bpca (Bayesian PCA) [189, 190] for MNAR values and MinProb [191] for MAR values. Differential analysis was performed using `limma` [71, 72] and `DEP` [70]. P values were adjusted using FDR correction by Benjamini-Hochberg [182]. The results of the differential interaction analysis were evaluated for functional enrichment by performing a gene set enrichment analysis (GSEA) using ClusterProfiler4 [192, 193] against the Gene Ontology (GO) Biological Process (BP) ontology [194, 195].

For the second approach, LFQ intensities were collapsed on GO BP terms by summing up all intensities of all identified proteins for each sample for each GO term. Then, for each GO term, a three-way ANOVA was performed with the main effects of mutation status (genetic context), condition and concentration

(culture context) and their interaction terms. The p-values of these ANOVAs were collectively corrected using correction by Holm [196]. After correction, significant terms ($p < 0.05$) were further analyzed using Tukey's Honest Significant Differences *post hoc* tests. P-values were collectively corrected using FDR correction by Benjamini-Hochberg [182].

Both approaches identify many GO terms that are significantly different ($p < 0.05$ after respective adjustment) between the groups. In order to gain an overview over the results, semantic similarity between the GO terms was determined using the methodology proposed by Schlicker et al. [197, 198]. Based on the resulting similarity matrix, GO terms were clustered using the binary cut algorithm [199]. The results were visualized as a heatmap with data from the analysis projected as additional heatmaps [199, 200]. All analysis in this part was performed using the R programming language ([69], `http://www.r-project.org/index.html`) and the tidyverse environment [83]. The scripts and output for this analysis are available on Zenodo (`https://zenodo.org/record/6896565`) (DOI: 10.5281/zenodo.6896565).

Visualization of AP-MS data in shiny app

The results from the functional analysis together with the filtered AP-MS data were put together in an R shiny dashboard, allowing the interactive exploration of our analysis and data [69, 83, 201, 202, 203, 204]. The R shiny app is available at `https://pjunk.shinyapps.io/kras_apms_vis/`, with the source code and underlying data files available at `https://github.com/PhilippJunk/kras_apms_vis`.

Assessment of phenotypic and metabolic parameters of Caco-2 cells

Caco-2 cells cultured in DMEM supplemented with 10% FBS were seeded at about 70% confluency in nine 12-well plates (CELLSTAR, Greiner Bio-One) to test 3 KRAS mutant status (WT, G12D and G12C) in three contexts (unstimulated, DMOG and IL-6). Twenty-four hours post-seeding cells were transfected with FLAG-KRASWT, FLAG-KRASG12D or flag-KRASG12C with

the protocol previously described. Then, 5 h post-transfection medium was changed and replaced with DMEM supplemented with 1% glutamine containing either 20 ng mL^{-1} DMOG, 20 ng mL^{-1} IL-6 or no stimulus (unstimulated). In addition, for cell transfected with KRASWT 15 ng mL^{-1} doxycycline was added for plasmid activation. Cell suspension samples were collected: once for all group during the seeding (24 h before transfection) then in triplicate for each group at 24 h, 48 h and 72 h post-transfection. Medium samples were collected: once during the context introduction (5 h post-transfection) then in triplicate for each group at 24 h, 48 h and 72 h post-transfection. Cell suspension samples were used for cell counting using Scepter^{TM} 2.0 Automated Cell Counter with 60 µm sensors (Merck Millipore), for cell viability and cellular ATP assessments using CellTiter-Glo®Luminescent Cell Viability Assay (Promega) and for Western-Blots of FLAG-KRAS (Sigma-Aldrich, Anti-FLAG®M2, F3165, mouse / monoclonal, 1:1000 dilution) normalized with β-actin (Cell Signaling, #4970, rabbit / monoclonal, 1:3000 dilution) using protocol previously described. For the Western Blot, after cell lysis, for each plasmid and at each time, the 3 context replicates were pooled to obtain enough protein to prepare 40 µL loading solution at 0.25 µg mL^{-1} (e.g., for WT at 24 h, the 3 replicates are 1 Unstim, 1 DMOG and 1 IL-6 samples). Medium were used for assessment of glucose uptake and lactate release using, respectively, Glucose-Glo^{TM} and Lactate-Glo^{TM} assays (Promega).

Statistical analysis

All data expressed as average $\pm$ standard deviations (SD), with SD represented by error bars. Statistical comparisons between two groups (typically treated group against control samples) were performed using a t-test. The average value and SD were calculated from at least three biological experiments. All tests were performed with a p-value of 0.05 using GraphPad Prism 9 software.

Network reconstruction and random walks analysis of AP-MS data

The starting point of the network are the 56 potential effectors of KRAS [30]. Then, beginning from these effectors, STRING (version 11.5) was used to construct the network [17]. All nodes which had a shortest path of 4 or less to these effectors were included, while filtering out edges with a STRING confidence score of less than 0.7. For KRAS, only edges towards the effectors were included in the network. Apart from the KRAS-effector edges, all interactions in the network are considered undirected. The final network consists of 15 062 nodes and 493 838 edges.

Using the network, targeted random walks were performed starting from KRAS, in the following called source, for each target protein in each condition of interest. For a predetermined number of steps, based on the current node, one of the connected nodes is randomly chosen. For each random walk, there is always only one target protein. As soon as the target protein is reached, or a certain number of iterations has been exceeded, the walk ends. The random walks are biased towards proteins found in the interactome of a certain condition. This is facilitated by favoring nodes found in the interactome by a factor of 20 over nodes not found the interactome of the specific condition. The actual probability depends on the number of connecting nodes.

$$P(in\ APMS) = 20 * P(not\ in\ APMS)$$

The number of iterations for the random walk for each target is dynamically calculated based on the length of the shortest path between source and target.

$$walklen = shortestpath + 2$$

Finally, the number of random walks, limited by runtime and memory, was set to 100 000 000 for each target for each condition. The code for the random walks was written in python using `numpy`, `scipy`, `pandas`, `numba` and `crsgraph` [102, 205, 206, 207, 208, 209]. The script used to run this analysis is available on Zenodo (`https://zenodo.org/record/6896565`) (DOI: 10.5281/zenodo.6896565).

Analysis of the random walks was performed by filtering out any paths that were found less than 100/100 000 000 walks and selecting the top 10 identified paths for each target by frequency. Paths were decomposed into a sequence of edges, and all edges for one condition were concatenated to generate a condition-specific network of information flow from KRAS to all proteins associated with a specific GO term. Networks were visualized, and the effector layer of each network was extracted and visualized together. All analysis and visualization were performed using R, in particular the packages `dplyr`, `tidyr`, `stringr`, `purrr`, `furrr`, `ggplot` and `ggraph` [184, 185, 186, 188, 210, 82, 211].

5.2.6 Author contributions

Methodology: CT, PJ, TS, SC, GO, KW, CK
Investigation: CT, PJ, TS, SC, GO, KW, CK
Writing – original draft: CT, TS, PJ, CK
Writing – review & editing: PJ, TS, SC, GO, KW

5.2.7 Acknowledgements

The authors would like to thank all members of the Kiel lab for discussions and critical reading of the manuscript. This work received funding from Science Foundation Ireland grant 16/FRL/3886 (CK) and from the Comprehensive Molecular Analytical Platform (CMAP) under The SFI Research Infrastructure Programme 18/RI/5702 (KW).

5.2.8 Supplement

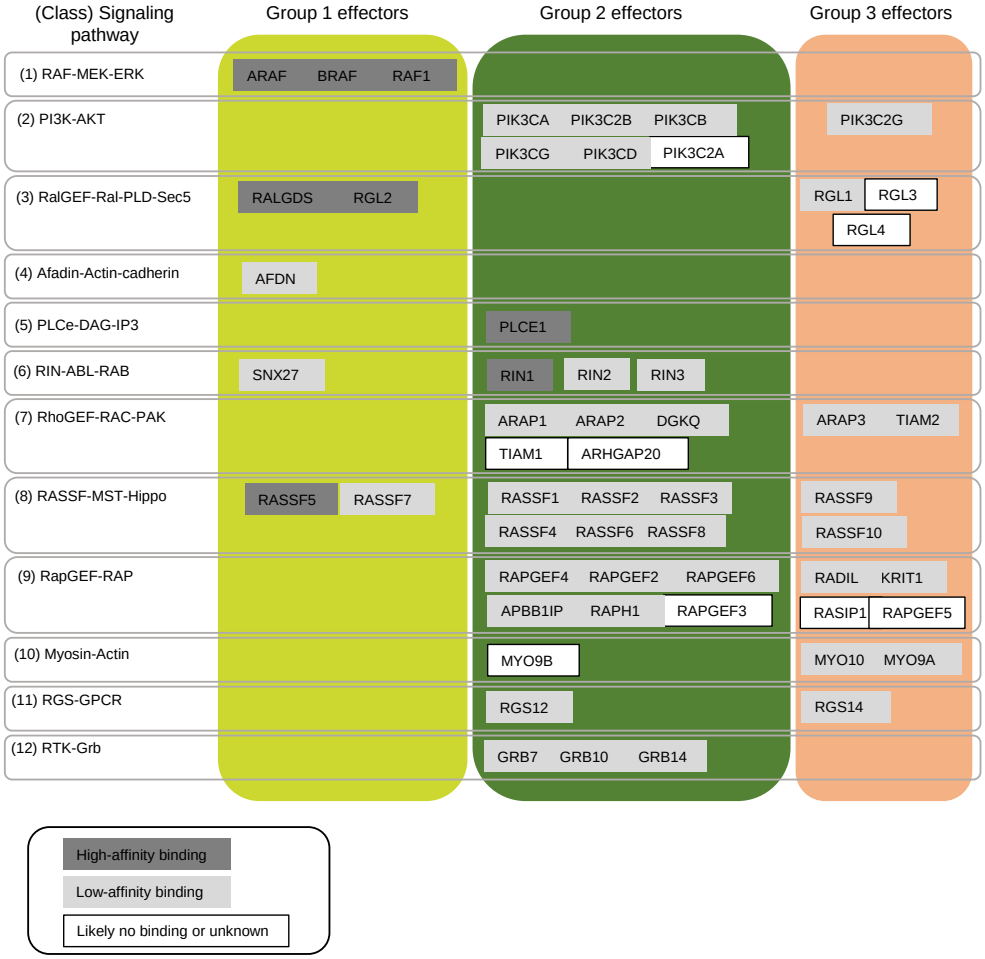

Figure 5.S1 Classification of 56 effectors into groups based on their requirement of additional domains for efficient binding to Ras-GTP. High-affinity effectors are defined by K_d values of $\leq$1 µM, and low- affinity binders by K_d values of >1 µM) (reviewed in [29] & [136]). Effectors groups (1-3) are based on predictions by [15]. Group 1 effectors form significant complexes with Ras-GTP using their RBD alone in most tissues. Group 2 effectors form significant complexes with Ras-GTP only with additional domains recruited to the PM. Group 3 effectors are predicted to be never in significant complex with Ras-GTP (in the 29 human tissues analysed) and are likely no true Ras effectors.

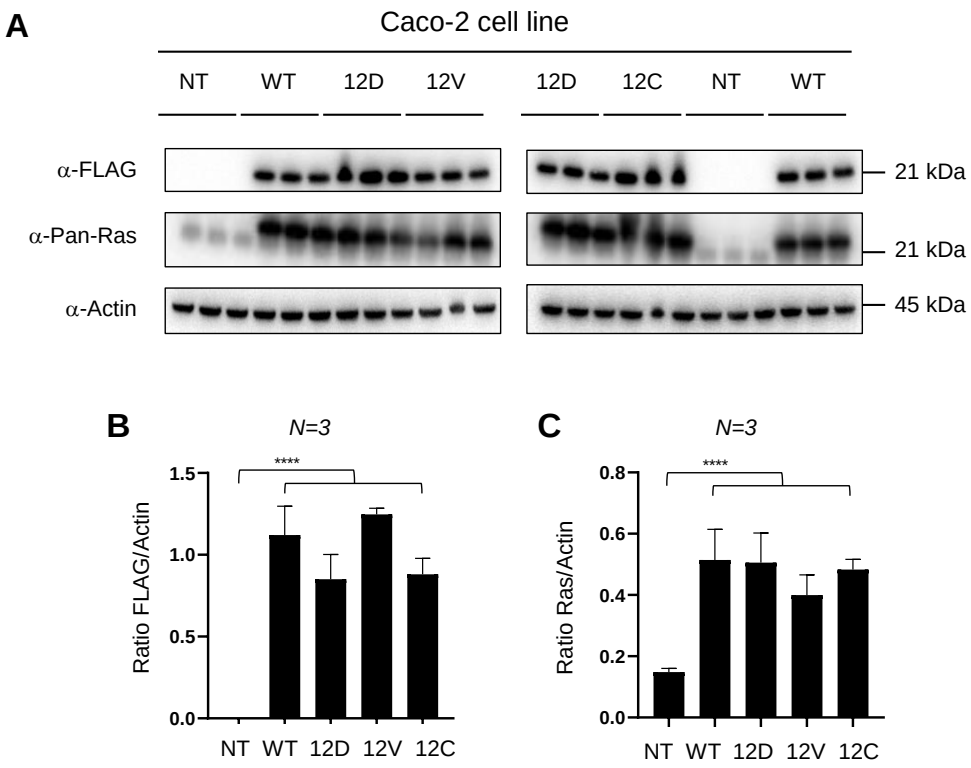

Figure 5.S2 Exogenous expression of FLAG-KRAS WT and mutant variants. (A) Results of Western blot analysis of FLAG-KRAS WT and mutant proteins using anti-FLAG, anti-Pan-Ras, and anti-Actin antibodies. (B) Quantification of exogenous KRAS normalized by actin as the loading control. (C) Quantification of the sum of endogenous Pan-Ras (= HRAS + NRAS + KRAS) and exogenous KRAS normalized by actin as the loading control. The KRAS WT biological triplicates were supplemented with 10 ng mL^{-1} of doxycycline after transfection. All samples were harvested at 24 h post-transfection. The ratios of FLAG/Actin and Ras/Actin analyses were performed with blots from N=3 independent experiments with ImageJ and analysed with two-way ANOVA.

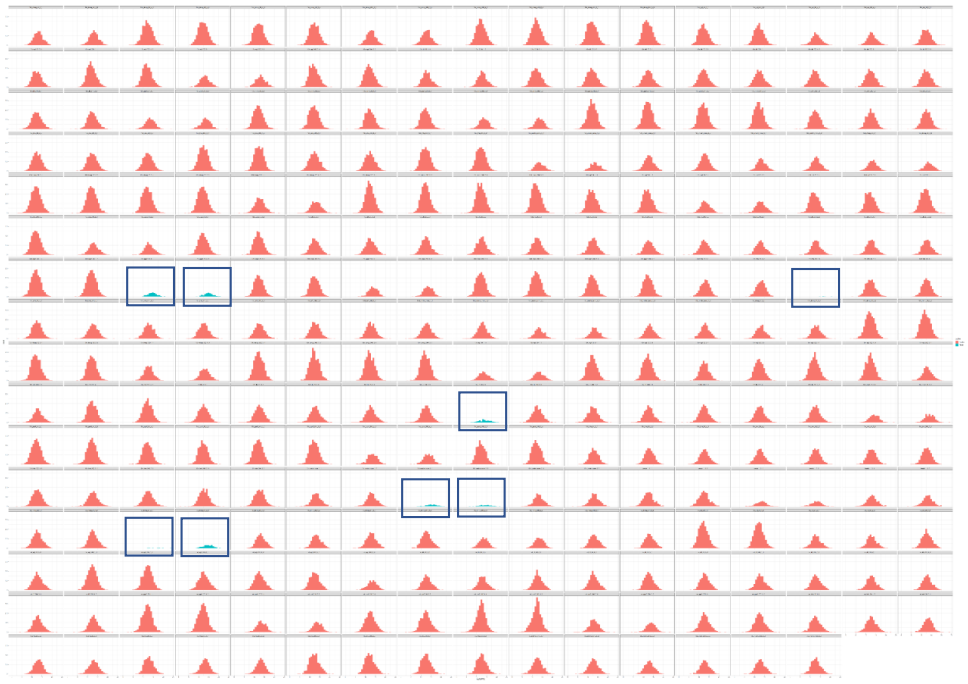

Figure 5.S3 Outlier analysis of AP-MS data. Each result of a MS analysis (biological or technical replicate) is represented as a histogram where the x-axis shows the LOG2 LFQ intensity of each protein and the y-axis shows the count (number of proteins). Histograms in red refer to datasets that are kept for further analyses. Histograms in blue and with boxes refer to outliers datasets that were filtered out (datasets: `12d_tnfa_200_1_2`; `12d_tnfa_200_2_1`; `12v_dmog_20_1_2`; `12v_pge2_200_3_1`; `wt_dmog_200_1_2`; `wt_dmog_200_2_1`; `wt_egf_200_1_2`; `wt_egf_200_2_1`).

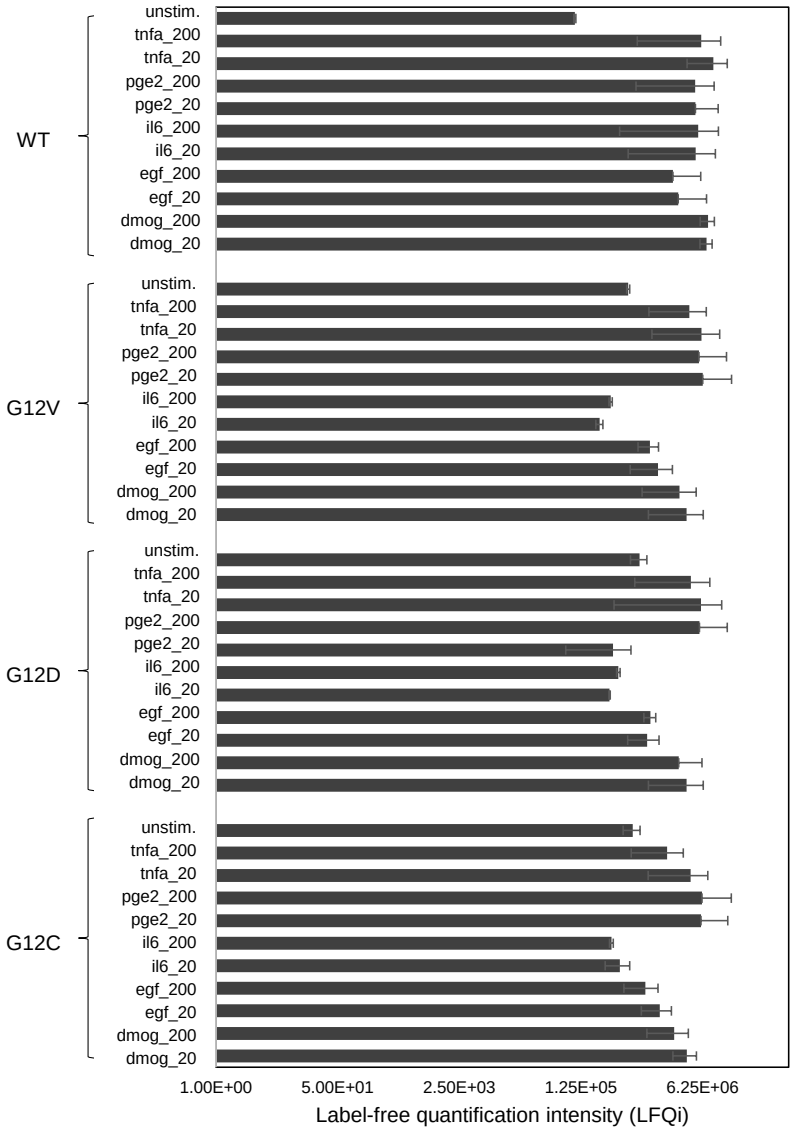

Figure 5.S4 KRAS protein abundances across the whole AP-MS dataset. The histogram represents the log10-fold change of the LFQ intensity of the WT and mutant KRAS proteins in the transfected Caco-2 cells with either WT, G12V, G12D, and G12C KRAS flagged proteins unstimulated (unstim) or stimulated with different growth conditions (i.e., DMOG, EGF, IL-6, PGE2, and TNF-α) at two concentrations (20 and 200 ng mL^{-1}). The average LFQ intensities are shown with error bars from the 3 biological replicates.

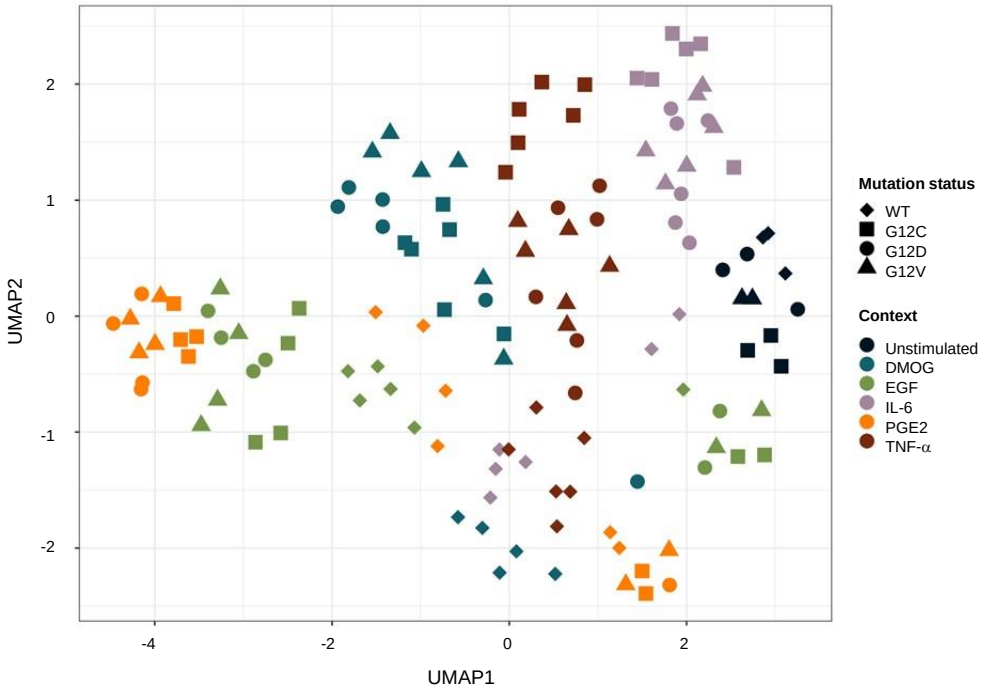

Figure 5.S5 UMAP analysis performed on LFQ intensity and executed with MS log2-transformed data after filtering on the whole AP-MS dataset. Colors indicate the different growth conditions, i.e., DMOG, EGF, IL-6, PGE2, TNF-α and unstimulated (unstim), and shapes indicate the concentration of the conditions (none, 20 ng mL^{-1} and 2200 ng mL^{-1}).

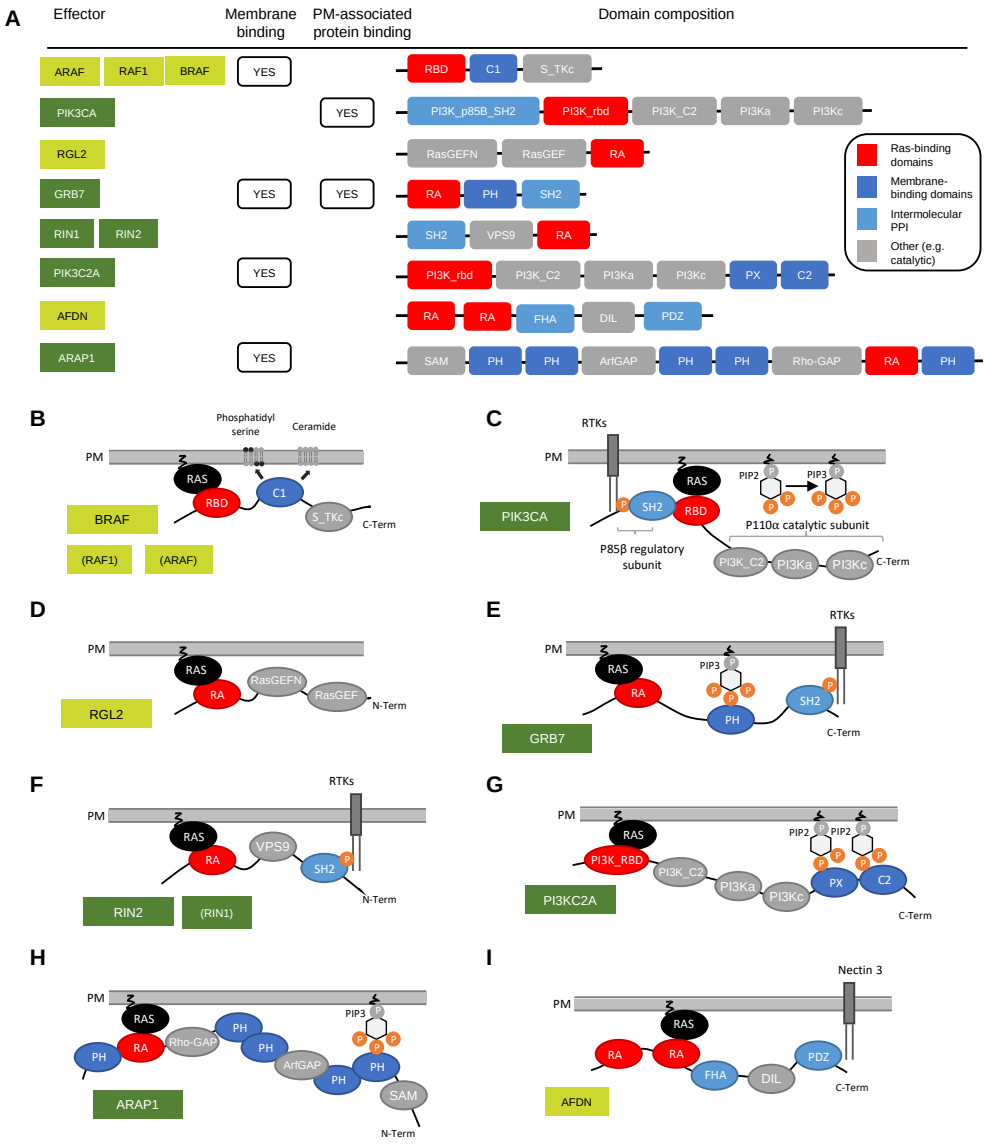

Figure 5.S6 Domain composition and capabilities of PM recruitment of 11 effectors identified in the AP-MS experiments. (A) Domain compositions of eleven effectors (adapted from [29]). Domain composition and ways of PM recruitment for BRAF, RAF1 and ARAF (B), PIK3CA (C), RGL2 (D), GRB7(E), RIN2 and RIN1 (F), PI3KC2A (G), ARAP1 (H), and AFDN (I).

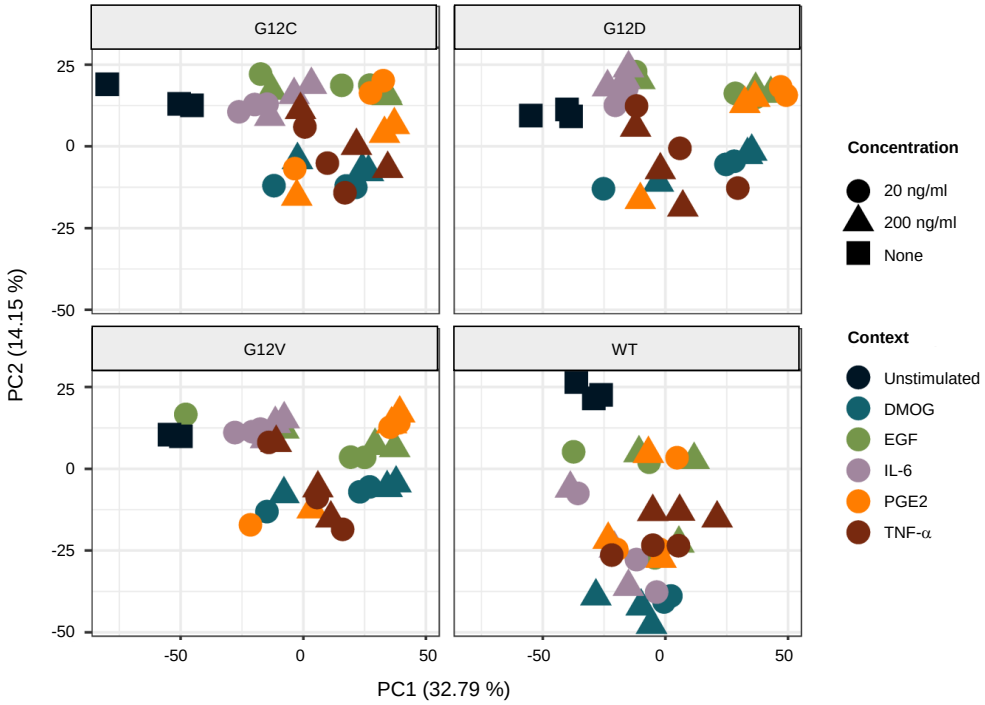

Figure 5.S7 Principle component (PC) analysis after transforming the data set by summing up LFQ intensities for each GO biological process. Facets indicate the different genetic context (WT, G12C, G12D, G12V), Colors indicate the different growth conditions (DMOG, EGF, IL-6, PGE2, TNF-α and unstimulated) and shapes indicate the concentration of the conditions (none, 20 ng mL^{-1} and 2200 ng mL^{-1}).

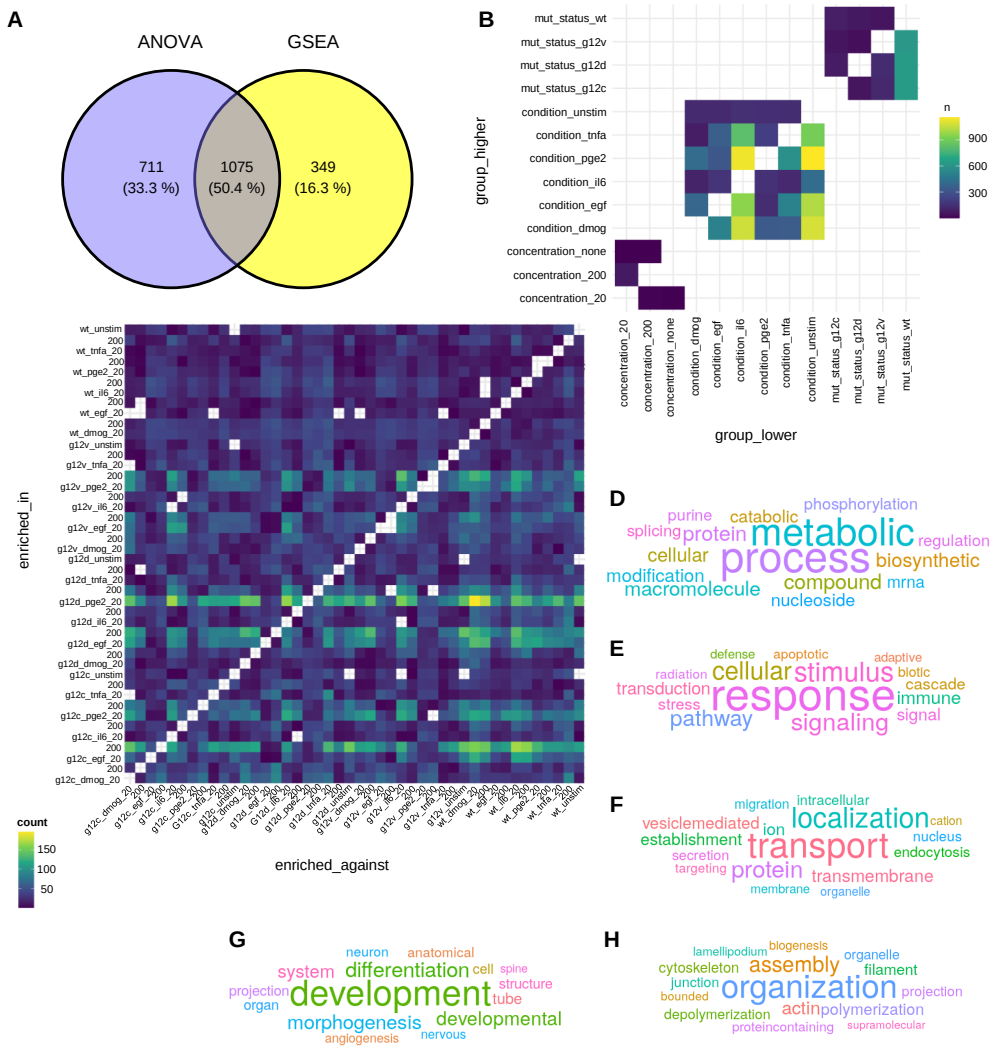

Figure 5.S8 Functional differences in the KRAS interactome. (A) Venn diagram showing the overlap of differentially identified GO terms between the two approaches. (B) Heatmap showing the number of significant (Adjusted p-value <0.05) contrasts for the main effects of the ANOVA analysis. (C) Heatmap showing the number of significantly (Adjusted p-value <0.05) enriched GO terms for the GSEA analysis between all samples. (D-H). Word clouds of the biggest semantic clusters from the GO term semantic similarity analysis (see Fig. 5.3). Visualized are cluster 1 (n=672), 2 (n=336), 3 (n=315), 4 (n=258) and 5 (n=221) in panels D, E, F, G, H, respectively.

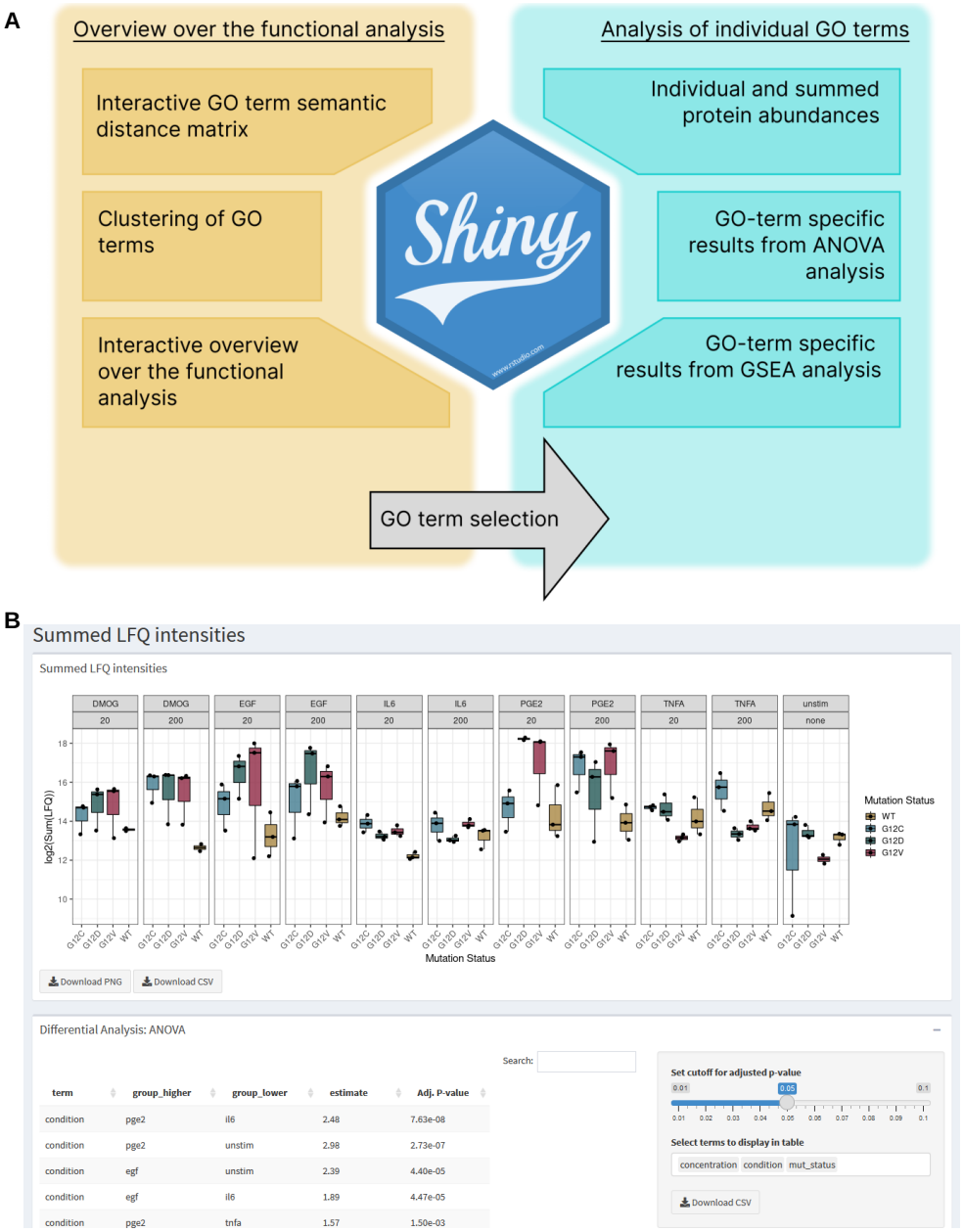

Figure 5.S9 Visualization of functional analysis of AP-MS experiments. (A) Overview and workflow of functionalities in the Shiny app. (B) Screenshot of data visualization using the Shiny app.

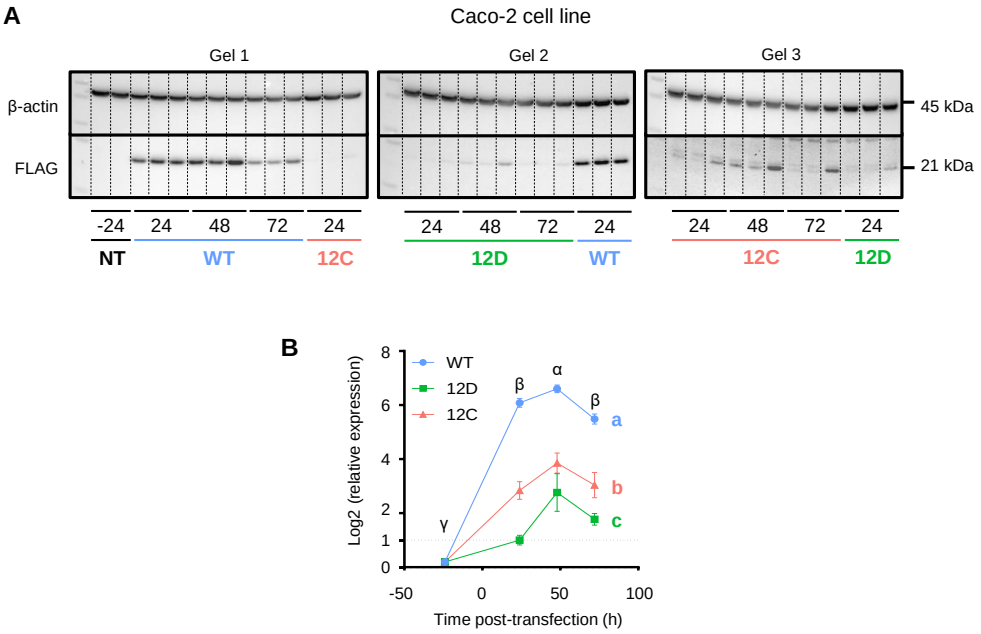

Figure 5.S10 Time course of exogenous expression of FLAG-KRAS WT and mutant during the Caco-2 phenotypical assessment experiment. (A) Results of Western Blot analysis of FLAG-KRAS WT and mutant using anti-FLAG and anti-Actin antibodies. For each plasmid at each time the protein sample come from a pool of the 3 context samples (1 Unstim, 1 DMOG and 1 IL-6 sample). (B) Relative quantification of FLAG-KRAS using the Gel Analyser tool in ImageJ. FLAG-KRAS expression normalized by β-actin as loading control have been normalized between gels using the average values of each plasmid at 24h. Values are mean ± SEM of log2 transformed and normalized FLAG expression relative to G12D at 24 h, N=3. Data were analysed with 2-ways ANOVA followed by Tukey's *post-hoc* test. Different letters represent significant differences ($p < 0.05$) between "mutant status" main effect and different Greek letters represent significant differences between different "time point" main effect.

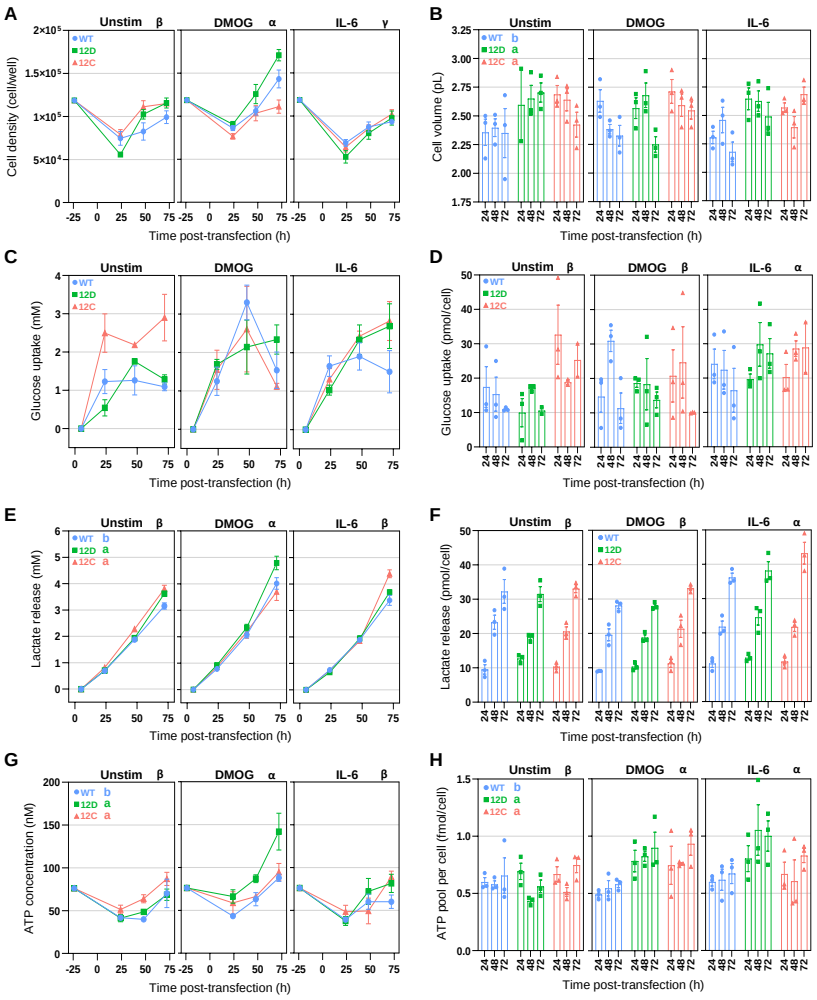

Figure 5.S11 Effect of culture and genetic contexts on phenotypic and metabolic parameters of Caco-2 cells. (A) cell density, (B) cell volume, (C) glucose uptake, (D) glucose uptake per cell, (E) lactate release, (F) lactate release per cell, (G) ATP concentration, (H) ATP pool per cell. Glucose uptake, lactate release and ATP pool were normalized per cell by dividing raw values by cell concentration (cellmL^{-1}). Values are mean ± SEM and N=3 per group. Data were analysed by a 3-way ANOVA with "culture context", "genetic context" and "time" as main effect. ANOVAs were followed by Tukey's *post-hoc* tests. Different letters represent significant differences ($p < 0.05$) between levels of the "genetic context" effect. Different Greek letters represents significant differences between levels of the "culture context" effect.

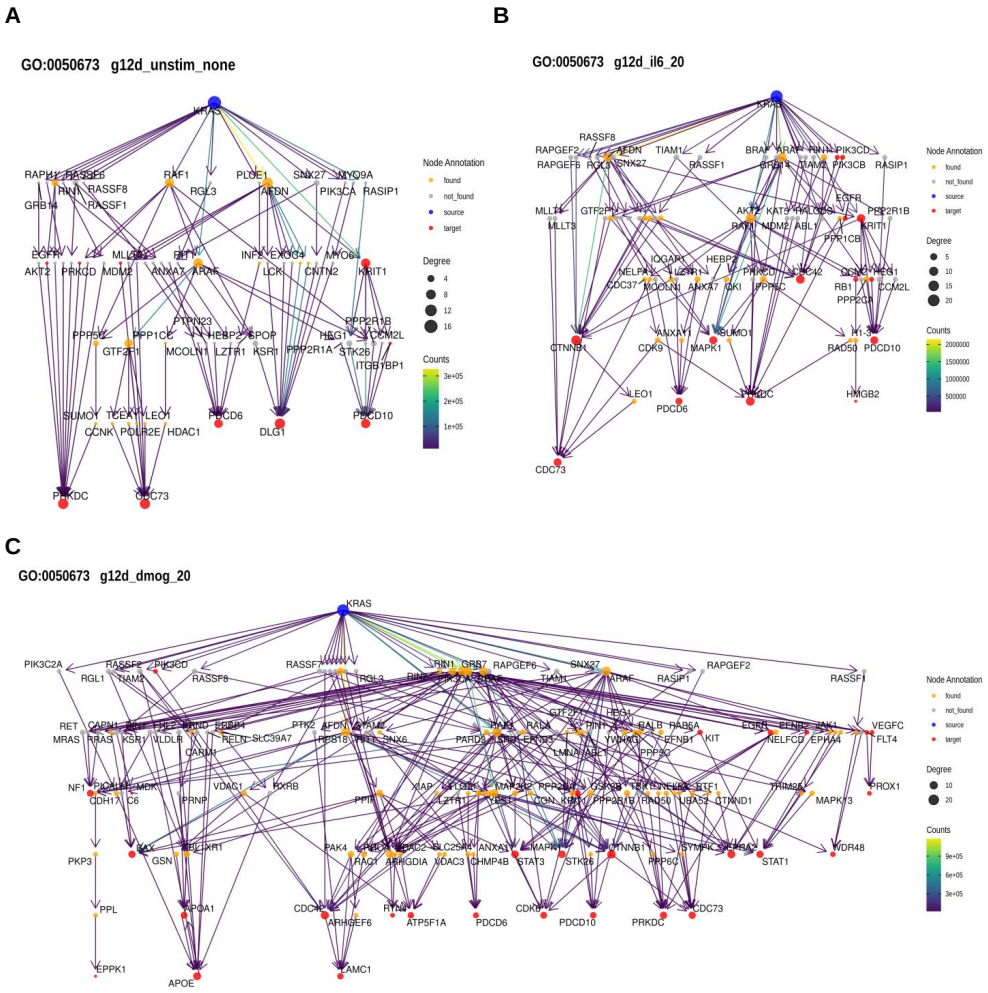

Figure 5.S12 Examples for random walk analyses linked to GO term “Epithelial cell proliferation” (GO:0050673). (A) AP-MS data for KRAS G12D in unstimulated culture context. (B) AP-MS data for KRAS G12D in unstimulated culture context. (C) AP-MS data for KRAS G12D in IL-6 culture context. (B) AP-MS data for KRAS G12D in IL-6 culture context.

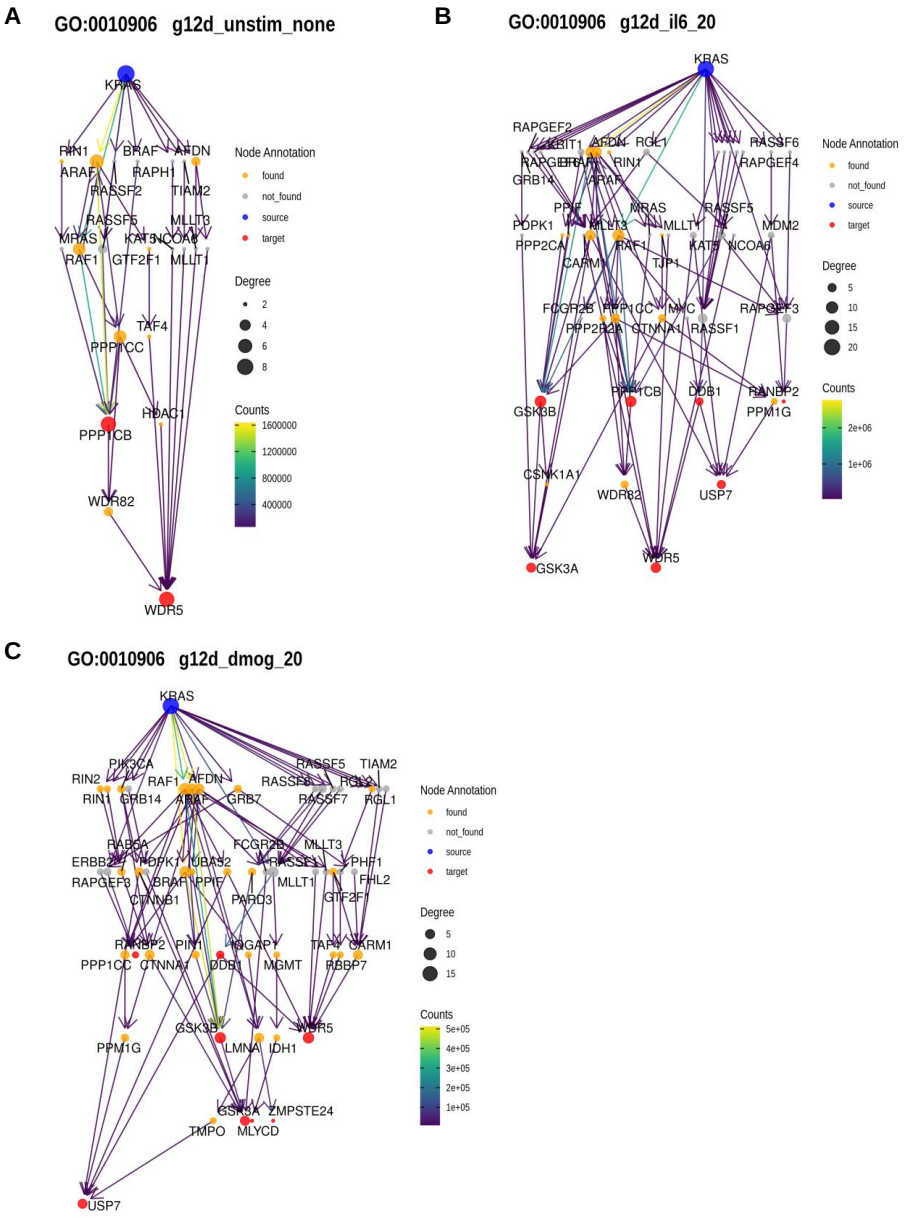

Figure 5.S13 Examples for random walk analyses linked to GO term “Regulation of glucose metabolic process” (GO:0010906). (A) AP-MS data for KRAS G12D in unstimulated culture context. (B) AP-MS data for KRAS G12D in unstimulated culture context. (C) AP-MS data for KRAS G12D in IL-6 culture context. (B) AP-MS data for KRAS G12D in IL-6 culture context.

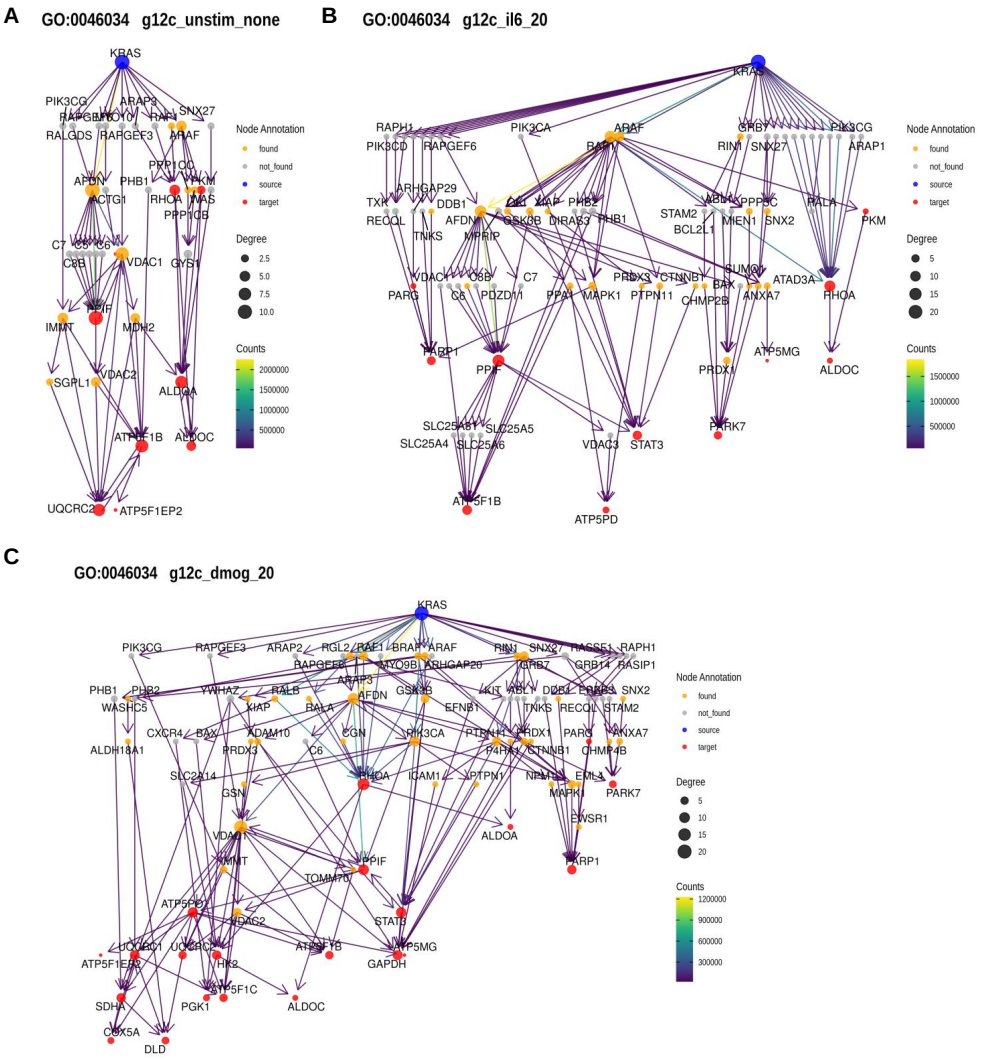

Figure 5.S14 Examples for random walk analyses linked to GO term “ATP metabolic process” (GO:0046034). (A) AP-MS data for KRAS G12D in unstimulated culture context. (B) AP-MS data for KRAS G12D in unstimulated culture context. (C) AP-MS data for KRAS G12D in IL-6 culture context. (B) AP-MS data for KRAS G12D in IL-6 culture context.

Table 5.S1 List of context-specific significant pairwise comparison of the sum of LFQ intensity of proteins in the AP-MS from six GO terms of interest (GO:0050673 Epithelial cell proliferation; GO:0050679 Positive regulation of cell population proliferation; GO:0006006 Glucose metabolic process; GO:0010906 Regulation of glucose metabolic process; GO:0046034 APT metabolic process; GO:1903578 Regulation of ATP metabolic process)(see Fig. 5.5). For each GO term the sum of LFQ intensity was analyse with a three-way ANOVA followed by Tukey *post-hoc* tests (see shiny app). Only pairwise comparisons of "culture" and "genetic" context effects and their two-way interaction with a p-value inferior to 0.1 are displayed. For the Genetic:Culture contexts interaction, only comparison of groups with one term in common are displayed. The six GO terms were selected for comparison to phenotypic parameters related to proliferation, glucose metabolism and ATP metabolism.

GO term	Pairwise Comparison			Statistics	
ID	Context effect	Group higher	Group lower	LS mean difference	p-adjusted
GO:0050673	Culture	DMOG	IL-6	2.227	4.36×10^{-9}
		DMOG	Unstim	2.891	4.08×10^{-9}
GO:0050679	Culture	DMOG	IL-6	2.682	3.46×10^{-7}
		DMOG	Unstim	5.347	4.07×10^{-9}
		IL-6	Unstim	2.665	2.66×10^{-4}
GO:0006006	Culture	DMOG	IL-6	1.116	3.16×10^{-4}
		DMOG	Unstim	2.738	4.07×10^{-9}
		IL-6	Unstim	1.622	1.54×10^{-5}
GO:0010906	Culture	DMOG	IL-6	0.781	4.14×10^{-4}
		DMOG	Unstim	2.905	4.07×10^{-9}
		IL-6	Unstim	2.124	4.07×10^{-9}
	Genetic	G12C	WT	0.846	1.18×10^{-6}
		G12D	WT	0.763	1.50×10^{-5}
	Genetic: Culture	G12C:DMOG	WT:DMOG	1.603	6.28×10^{-3}
		G12D:DMOG	WT:DMOG	1.513	1.53×10^{-2}
		G12C:IL-6	WT:IL-6	1.365	3.30×10^{-2}
		G12D:DMOG	G12D:Unstim	3.386	4.59×10^{-9}
		G12D:IL-6	G12D:Unstim	2.839	4.45×10^{-7}
		G12C:IL-6	G12C:Unstim	2.697	1.98×10^{-6}
		G12C:DMOG	G12C:Unstim	3.305	6.06×10^{-9}
GO:0046034	Culture	DMOG	IL-6	1.312	1.79×10^{-7}
		DMOG	Unstim	0.790	5.78×10^{-2}
GO:1903578	Culture	DMOG	IL-6	1.335	1.50×10^{-2}
		DMOG	Unstim	1.831	4.84×10^{-3}

5.3 Discussion

In this work, we explored the interactome of KRAS, with different mutation status and under different treatments. We mapped protein-protein interactions using interaction proteomics and analysed functional differences between different conditions. Finally, we analysed how signals could propagate from KRAS to cause different functional effects.

As highlighted previously, my contribution to this paper has been the bioinformatic analysis (a detailed description can be found in the preamble on page vii). Nonetheless, I want to briefly touch on the experimental side of the project as well. In this work, the aim was the analysis of KRAS signalling networks in different conditions that are to a certain degree resembling the (patho-) physiological context in the colon epithelium. For this, several decisions on the level of the experimental system were made: A) the choice of cell line; B) the choice of how to include oncogenic variants of KRAS in the system; and C) the choice of treatments to stimulate the cells with.

Protein-protein interactions are usually measured in non-physiological contexts and cells, such as HEK293T cells [170]. However, it has been shown that protein-protein interactions rewire between different cell types and differences in the environment of cells [212, 213]. In order to investigate protein-protein interaction networks in a model close to early adenocarcinoma, the cell line of choice was Caco-2 cells. These cells are an interesting system for multiple reasons. Caco-2 cells are mutated in APC, CTNNB1 (β-catenin), SMAD4, and TP53, but not in KRAS [62, 63, 64]. These mutations match what is known as the Vogelstein sequence [148], establishing them as a model of early colorecal adenocarcinoma. From the perspective of physiology, Caco-2 cells are able to spontaneously differentiate into an epithelium, both in 2D and 3D culture. Additionally, Caco-2 cells are well established for studies of intestinal epithelial permeability, indicating that they are a relevant model for colon epithelium. However, the individual cells can be diverse in shape and size, introducing biological variablility into the experiments. As I am not an expert on different cell lines for colon cancer and epithelium, I am going with the judgement of my wet lab colleagues that, while Caco-2 cells are not the ideal system, they are

an interesting and relevant system to study KRAS interactions in the (patho-) physiological context of early adenocarcinoma.

Moving on to the choice of how to introduce oncogenic variants of KRAS. Originally, the idea was to use CRISPR/Cas9 to create point mutations for the oncogenic variants of interest and work with endogenous KRAS expression. However, this was not successful for multiple reasons. Caco-2 cells are slowly growing cells, in particular when growing in isolation, which is unfortunately part of the protocol for generating genotypically identical clones. Not many clones survived, and those that did were not of the desired genotype. Secondly, a lot of the antibodies against KRAS were not strong enough binders to perform Co-IP experiments on endogenous KRAS consistently. For this reason, the decision was made to pivot to an exogenous expression of flag-tagged KRAS. More details on these approaches and how they were performed can be found in the PhD thesis of my colleague who had been performing the experiments [214].

Regarding the choice of treatments to stimulate the cells with, these were chosen based on a detailed literature analysis of the extracellular signals KRAS is integrating in the physiological and oncogenic intestinal epithelium. The treatments were chosen to mimic relevant influences such as the tumour microenvironment, growth signalling and hypoxia. These treatments are clearly artificial simplifications of what is actually happening to cancer cells in their environment, however this approach gives us the option to investigate their influence on the KRAS interactome in separation. However, the artificial nature of these treatments is a limiting factor for the experiments. For example, the treatment of the cells with DMOG as a proxy for hypoxia only partially mimics the hypoxic response of cells. In detail, DMOG is an inhibitor of prolyl hydroxylases 1-3 (PHD1-3), a family of enzymes which in normoxic conditions lead to the degradation of HIF-α [65]. As their enzymatic activity is dependent on O_2, under hypoxic conditions, HIF-α is stabilized and can act as a transcription factor. However, it is becoming increasingly clearer that there are also PHD/HIF independent adaptation to low oxygen levels which play an important role in the hypoxia response [66, 215, 216]. The treatment of the cells with DMOG can therefore not be interpreted as an equivalent of low O2 exposure, which

might have been the better experimental setup to use for these experiments. Similar limitations apply to the other treatments, i.e., the interaction of immune cells in the tumour microenvironment cannot be accurately mimicked by the addition of one or multiple cytokines and chemokines.

In the context of DMOG, it has been brought to my attention that the concentration used in the publication is lower than what is typically used in the literature. In the publication, we use 20 and 200 ng mL^{-1}, which is 0.114 and 1.14 µM, respectively. Common values in the literature are around 100 µM to 2 mM [217]. Interestingly, the DMOG treated samples are showing one of the strongest differences compared to the untreated samples, even at these low concentrations. To investigate whether the observed changes in the interactome are due to a stabilization of HIF-alpha, it would have been good to check for stabilization of HIF-α with a Western blot at different concentrations of DMOG.

Next, I want to discuss the bioinformatics analysis, which I have performed. The analysis of the proteomics data was straightforward, using well established software and resources. More interestingly, I have tried to provide an interactive interface to the data analysis by designing and publishing an R shiny web app. Designing good user interfaces is hard, and I am not an expert in it, so I think there is a lot to improve in terms of user interface and experience. However, I think the concept of making data interactive is something that could be interesting for more biological projects, although additional time and skills are required for the design of these platforms.

Finally, I tried to approach the question about how the signal propagates from KRAS to certain functional submodules through the network in this publication by performing random walks over an interaction network, which was biased by the presence or absence of specific proteins in the AP/MS data. This is to the best of my knowledge a novel approach, and there are certainly some aspects that can be improved on. In particular, a better statistical framework to analyse and normalize the results from the random walks would be interesting. Ideas that come to mind, but would need to be tested, are comparisons against random networks or random biases. Also, currently the path length is a major

bias in determining the counts for specific paths through the networks, which needs to be considered and potentially corrected.

In terms on limitations of the approach, there are two major ones. Firstly, the approach is limited by the data used for the interaction network. In this publication, we used the STRING database filtered for high confidence interactors. However, different databases for protein interactions come from different background and with their specific biases. Secondly, proteomic approaches are well known to have problems with noise and missing values. Since these are used as inputs for the random walks, these problems propagate to this method as well.

Bibliography

[1] O. Wolkenhauer. Why model? *Frontiers in Physiology*, 5 JAN:21, 2014.

[2] P. Aloy and R. B. Russell.
Structural systems biology: modelling protein interactions.
Nature Reviews Molecular Cell Biology, 7(3):188–197, 2006.

[3] H. Kitano. Computational systems biology.
Nature, 420(6912):206–210, 2002.

[4] R. A. Hillmer. Systems biology for biologists.
PLOS Pathogens, 11(5):e1004786, 2015.

[5] P. Beltrao, C. Kiel, and L. Serrano. Structures in systems biology.
Current Opinion in Structural Biology, 17(3):378–384, 2007.

[6] S. Curry.
Structural biology: a century-long journey into an unseen world.
Interdisciplinary Science Reviews, 40(3):308–328, 2015.

[7] S. K. Burley et al. Protein data bank: the single global archive for 3d macromolecular structure data.
Nucleic Acids Research, 47:D520–D528, D1, 2019.

[8] T. Hameduh et al.
Homology modeling in the time of collective and artificial intelligence.
Computational and Structural Biotechnology Journal, 18:3494–3506, 2020.

[9] J. Jumper et al.
Highly accurate protein structure prediction with alphafold.
Nature, 596(7873):583–589, 2021.

[10] M. AlQuraishi. Protein-structure prediction revolutionized. *Nature*, 596(7873):487–488, 2021.

[11] D. T. Eriksen, J. Lian, and H. Zhao. Protein design for pathway engineering. *Journal of Structural Biology*, 185(2):234–242, 2014.

[12] C. Kiel, E. Yus, and L. Serrano. Engineering signal transduction pathways. *Cell*, 140(1):33–47, 2010.

[13] B. N. Kholodenko et al. Quantification of short term signaling by the epidermal growth factor receptor. *Journal of Biological Chemistry*, 274(42):30169–30181, 1999.

[14] D. Fey et al. Signaling pathway models as biomarkers: patient-specific simulations of jnk activity predict the survival of neuroblastoma patients. *Science Signaling*, 8(408), 2015.

[15] S. Catozzi, M. Halasz, and C. Kiel. Predicted 'wiring landscape' of ras-effector interactions in 29 human tissues. *npj Systems Biology and Applications*, 7(1):1–15, 2021.

[16] M. Vidal, M. E. Cusick, and A. L. Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[17] D. Szklarczyk et al. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47:D607–D613, D1, 2019.

[18] L. F. Iglesias-Martinez, B. D. Kegel, and W. Kolch. Kboost: a new method to infer gene regulatory networks from gene expression data. *Scientific Reports*, 11(1):1–13, 2021.

[19] K. Luck et al. A reference map of the human binary protein interactome. *Nature*, 580(7803):402–408, 2020.

[20] M. E. Sardiu and M. P. Washburn. Building protein-protein interaction networks with proteomics and informatics tools. *Journal of Biological Chemistry*, 286(27):23645–23651, 2011.

[21] M. Kim et al. A protein interaction landscape of breast cancer. *Science*, 374(6563), 2021.

[22] Q. Zhong et al.
Edgetic perturbation models of human inherited disorders.
Molecular Systems Biology, 5(1):321, 2009.

[23] C. Kiel and L. Serrano. Structure-energy-based predictions and network modelling of rasopathy and cancer missense mutations. *Molecular Systems Biology*, 10(5):727, 2014.

[24] O. S. Rukhlenko et al.
Dissecting raf inhibitor resistance by structure-based modeling reveals ways to overcome oncogenic ras signaling.
Cell Systems, 7(2):161–179.e14, 2018.

[25] T. Ideker, T. Galitski, and L. Hood.
A new approach to decoding life: systems biology. *Annual Review of Genomics and Human Genetics*, 2(1):343–372, 2003.

[26] S. Lu et al. Ras conformational ensembles, allostery, and signaling. *Chemical Reviews*, 116(11):6607–6665, 2016.

[27] G. A. Hobbs, C. J. Der, and K. L. Rossman.
Ras isoforms and mutations in cancer at a glance.
Journal of Cell Science, 129(7):1287–1292, 2016.

[28] J. L. Bos, H. Rehmann, and A. Wittinghofer.
Gefs and gaps: critical elements in the control of small g proteins.
Cell, 129(5):865–877, 2007.

[29] C. Kiel, D. Matallanas, and W. Kolch.
The ins and outs of ras effector complexes.
Biomolecules, 11(2):236, 2021.

[30] V. I. Gaspar et al. Analysis of ras-effector interaction competition in large intestine and colorectal cancer context.
Small GTPases, 12(3):209–225, 2021.

[31] K. Rajalingam et al. Ras oncogenes and their downstream targets.
Biochimica et Biophysica Acta (BBA) - Molecular Cell Research, 1773(8):1177–1195, 2007.

[32] S. Singh and M. J. Smith.
Ras gtpase signalling to alternative effector pathways.
Biochemical Society Transactions, 48(5):2241–2252, 2020.

[33] D. K. Simanshu, D. V. Nissley, and F. McCormick.
Ras proteins and their regulators in human disease.
Cell, 170(1):17–33, 2017.

[34] C. Ternet and C. Kiel. Signaling pathways in intestinal homeostasis and colorectal cancer: kras at centre stage.
Cell Communication and Signaling, 19(1):1–22, 2021.

[35] M. Malumbres and M. Barbacid. Ras oncogenes: the first 30 years.
Nature Reviews Cancer, 3(6):459–465, 2003.

[36] A. M. D. Vos et al. Three-dimensional structure of an oncogene protein: catalytic domain of human c-h-ras p21.
Science, 239(4842):888–893, 1988.

[37] E. F. Pai et al. Structure of the guanine-nucleotide-binding domain of the ha-ras oncogene product p21 in the triphosphate conformation.
Nature, 341(6239):209–214, 1989.

[38] A. Erijman and J. M. Shifman.
Ras/effector interactions from structural and biophysical perspective.
Mini reviews in medicinal chemistry, 16(5):370–5, 2016.

[39] J. M. Ostrem et al. K-ras(g12c) inhibitors allosterically control gtp affinity and effector interactions.
Nature 2013 503:7477, 503:548–551, 7477, 2013.

[40] C. Kiel and L. Serrano. The ubiquitin domain superfold: structure-based sequence alignments and characterization of binding epitopes.
Journal of Molecular Biology, 355(4):821–844, 2006.

[41] M. J. Smith, B. G. Neel, and M. Ikura. Nmr-based functional profiling of rasopathies and oncogenic ras mutations. *Proceedings of the National Academy of Sciences*, 110(12):4574–9, 2013.

[42] M. J. Smith and M. Ikura. Integrated ras signaling defined by parallel nmr detection of effectors and regulators. *Nature Chemical Biology*, 10(3):223–230, 2014.

[43] J. C. Hunter et al. Biochemical and structural analysis of common cancer-associated kras mutations. *Molecular Cancer Research*, 13(9):1325–1335, 2015.

[44] M. V. Cespedes et al. K-ras asp12 mutant neither interacts with raf, nor signals through erk and is less tumorigenic than k-ras val12. *Carcinogenesis*, 27(11):2190–2200, 2006.

[45] N. T. Ihle et al. Effect of kras oncogene substitutions on protein behavior: implications for signaling and clinical outcome. *JNCI: Journal of the National Cancer Institute*, 104(3):228–239, 2012.

[46] B. Stolze et al. Comparative analysis of kras codon 12, 13, 18, 61 and 117 mutations using human mcf10a isogenic cell lines. *Scientific Reports*, 5(1):8535, 2015.

[47] C. Muñoz-Maldonado, Y. Zimmer, and M. Medová. A comparative analysis of individual ras mutations in cancer biology. *Frontiers in Oncology*, 9:1088, 2019.

[48] S. Lu et al. The structural basis of oncogenic mutations g12, g13 and q61 in small gtpase k-ras4b. *Scientific Reports*, 6(1):21949, 2016.

[49] C. Kiel et al. A genome-wide ras-effector interaction network. *Journal of Molecular Biology*, 370(5):1020–1032, 2007.

[50] C. Kiel and L. Serrano. Prediction of ras-effector interactions using position energy matrices. *Bioinformatics*, 23(17):2226–2230, 2007.

[51] K. Illergård, D. H. Ardell, and A. Elofsson.
Structure is three to ten times more conserved than sequence—a study of structural response in protein cores.
Proteins: Structure, Function, and Bioinformatics, 77(3):499–508, 2009.

[52] E. Brunk et al. Systems biology of the structural proteome.
BMC Systems Biology, 10(1):1–16, 2016.

[53] E. Brunk et al. Recon3d enables a three-dimensional view of gene variation in human metabolism.
Nature Biotechnology, 36(3):272–281, 2018.

[54] D. Abankwa and A. A. Gorfe. Mechanisms of ras membrane organization and signaling: ras rocks again.
Biomolecules, 10(11):1522, 2020.

[55] E. S. Ozdemir, A. M. Koester, and X. Nan. Ras multimers on the membrane: many ways for a heart-to-heart conversation.
Genes, 13(2):219, 2022.

[56] V. P. Mysore et al. A structural model of a ras–raf signalosome.
Nature Structural & Molecular Biology, 28(10):847–857, 2021.

[57] S. Catozzi et al. Reconstruction and analysis of a large-scale binary ras-effector signaling network.
Cell Communication and Signaling, 20(1):1–19, 2022.

[58] T. H. Tran et al. Kras interaction with raf1 ras-binding domain and cysteine-rich domain provides insights into ras-mediated raf activation.
Nature Communications, 12(1):1–16, 2021.

[59] M. Varadi et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models.
Nucleic Acids Research, 50:D439–D444, D1, 2022.

[60] D. Wang et al. A deep proteome and transcriptome abundance atlas of 29 healthy human tissues.
Molecular Systems Biology, 15(2):e8503, 2019.

[61] J. Fogh, W. C. Wright, and J. D. Loveless. Absence of hela cell contamination in 169 cell lines derived from human tumors. *JNCI: Journal of the National Cancer Institute*, 58(2):209–214, 1977.

[62] A. Bairoch. The cellosaurus, a cell-line knowledge resource. *Journal of biomolecular techniques : JBT*, 29:25–38, 2, 2018.

[63] M. Ilyas et al. Beta-catenin mutations in cell lines established from human colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 94:10330–10334, 19, 1997.

[64] D. Mouradov et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Research*, 74:3238–3247, 12, 2014.

[65] A. J. Majmundar, W. J. Wong, and M. C. Simon. Hypoxia-inducible factors and the response to hypoxic stress. *Molecular Cell*, 40:294–309, 2, 2010.

[66] L. Iommarini et al. Non-canonical mechanisms regulating hypoxia-inducible factor 1 alpha in cancer. *Frontiers in Oncology*, 7:286, NOV, 2017.

[67] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12):1367–1372, 2008.

[68] J. Cox et al. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed maxlfq. *Molecular and Cellular Proteomics*, 13(9):2513–2526, 2014.

[69] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022.

[70] X. Zhang et al. Proteome-wide identification of ubiquitin interactions using ubia-ms. *Nature Protocols*, 13(3):530–550, 2018.

[71] M. E. Ritchie et al. Limma powers differential expression analyses for rna-sequencing and microarray studies.
Nucleic Acids Research, 43(7):e47–e47, 2015.

[72] B. Phipson et al.
Robust hyperparameter estimation protects against hypervariable genes and improves power to detect differential expression.
The Annals of Applied Statistics, 10(2):946–963, 2016.

[73] Q. N. Van et al. Ras nanoclusters: dynamic signaling platforms amenable to therapeutic intervention. *Biomolecules*, 11(3):377, 2021.

[74] R. Guerois, J. E. Nielsen, and L. Serrano.
Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations.
Journal of Molecular Biology, 320(2):369–387, 2002.

[75] J. Schymkowitz et al. The foldx web server: an online force field.
Nucleic Acids Research, 33:W382–W388, suppl_2, 2005.

[76] C. Kiel, L. Serrano, and C. Herrmann.
A detailed thermodynamic analysis of ras/effector complex interfaces.
Journal of Molecular Biology, 340(5):1039–1058, 2004.

[77] C. Kiel and L. Serrano. Cell type-specific importance of ras-c-raf complex association rate constants for mapk signaling.
Science Signaling, 2(81), 2009.

[78] K. Stojanovski et al. Interaction dynamics determine signaling and output pathway responses. *Cell Reports*, 19(1):136–149, 2017.

[79] R. Qamra and S. R. Hubbard. Structural basis for the interaction of the adaptor protein grb14 with activated ras.
PLOS ONE, 8(8):e72473, 2013.

[80] D. Filchtinski et al. What makes ras an efficient molecular switch: a computational, biophysical, and structural study of ras-gdp interactions with mutants of raf.
Journal of Molecular Biology, 399(3):422–435, 2010.

[81] M. J. Smith et al. Evolution of af6-ras association and its implications in mixed-lineage leukemia. *Nature Communications*, 8(1):1–13, 2017.

[82] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[83] H. Wickham et al. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

[84] N. Nassar et al. Ras/rap effector specificity determined by charge reversal. *Nature Structural Biology*, 3(8):723–729, 1996.

[85] J. van Durme et al. A graphical interface for the foldx forcefield. *Bioinformatics*, 27(12):1711–1712, 2011.

[86] E. Krieger and G. Vriend. Yasara view—molecular graphics for all devices—from smartphones to workstations. *Bioinformatics*, 30(20):2981–2982, 2014.

[87] J. Delgado et al. Foldx 5.0: working with rna, small molecules and a new graphical interface. *Bioinformatics*, 35:4168–4169, 20, 2019.

[88] R. F. Alford et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13:3031–3048, 6, 2017.

[89] A. Šali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993.

[90] G. Q. Dong et al. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics*, 29(24):3158–3166, 2013.

[91] P. Benkert, S. C. Tosatto, and D. Schomburg. Qmean: a comprehensive scoring function for model quality assessment. *Proteins: Structure, Function and Genetics*, 71(1):261–277, 2008.

[92] P. Benkert, M. Biasini, and T. Schwede. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*, 27(3):343–350, 2011.

[93] G. Studer et al.
Qmeandisco-distance constraints applied on model quality estimation.
Bioinformatics, 36(6):1765–1771, 2020.

[94] V. B. Chen et al. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66(1):12–21, 2010.

[95] C. J. Williams et al. Molprobity: more and better reference data for improved all-atom structure validation.
Protein Science, 27(1):293–315, 2018.

[96] B. Webb and A. Sali.
Comparative protein structure modeling using modeller.
Current Protocols in Bioinformatics, 54(1):5.6.1–5.6.37, 2016.

[97] G. Janson et al. Revisiting the “satisfaction of spatial restraints” approach of modeller for protein homology modeling. *PLoS Computational Biology*, 15(12):e1007219, 2019. B. L. de Groot, editor.

[98] M. Biasini et al. Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69(5):701–709, 2013.

[99] G. Studer et al. Promod3—a versatile homology modelling toolbox.
PLOS Computational Biology, 17(1):e1008667, 2021.

[100] M. Shen and A. Sali.
Statistical potential for assessment and prediction of protein structures.
Protein Science, 15(11):2507–2524, 2006.

[101] A. Waterhouse et al.
Swiss-model: homology modelling of protein structures and complexes.
Nucleic Acids Research, 46:W296–W303, W1, 2018.

[102] G. V. Rossum and F. L. Drake. *Python 3 Reference Manual*.
CreateSpace, 2009.

[103] Y. Rose et al. Rcsb protein data bank: architectural advances towards integrated searching and efficient access to macromolecular structure data from the pdb archive.
Journal of Molecular Biology, 433(11):166704, 2021.

[104] M. Steinegger and J. Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.
Nature Biotechnology, 35(11):1026–1028, 2017.

[105] F. Sievers et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega.
Molecular Systems Biology, 7(1):539, 2011.

[106] F. Sievers and D. G. Higgins. Clustal omega for making accurate alignments of many protein sequences.
Protein Science, 27(1):135–145, 2018.

[107] M. Steinegger et al. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1):1–15, 2019.

[108] R. Gasper and F. Wittinghofer.
The ras switch in structural and historical perspective.
Biological Chemistry, 401(1):143–163, 2019.

[109] E. F. Pai et al.
Refined crystal structure of the triphosphate conformation of h-ras p21 at 1.35 a resolution: implications for the mechanism of gtp hydrolysis.
The EMBO Journal, 9(8):2351–2359, 1990.

[110] C. Kiel et al. Integration of protein abundance and structure data reveals competition in the erbb signaling network.
Science Signaling, 6(306), 2013.

[111] B. N. Kholodenko, J. B. Hoek, and H. V. Westerhoff. Why cytoplasmic signalling proteins should be recruited to cell membranes.
Trends in Cell Biology, 10(5):173–178, 2000.

[112] R. C. Gimple and X. Wang.
Ras: striking at the core of the oncogenic circuitry.
Frontiers in Oncology, 9:965, 2019.

[113] A. R. Moore et al. Ras-targeted therapies: is the undruggable drugged? *Nature Reviews Drug Discovery*, 19(8):533–552, 2020.

[114] A. G. Stephen et al. Dragging ras back in the ring. *Cancer Cell*, 25(3):272–281, 2014.

[115] P. Junk and C. Kiel. Engineering of biological pathways: complex formation and signal transduction. *Methods in Molecular Biology*, 2315:59–70, 2021.

[116] P. Junk and C. Kiel. Homelette: a unified interface to homology modelling software. *Bioinformatics*, 38(6):1749–1751, 2022.

[117] I. Jarmoskaite et al. How to measure and evaluate binding affinities. *eLife*, 9:1–34, 2020.

[118] I. Humphreys et al. Computed structures of core eukaryotic protein complexes. *Science*, 374(6573), 2021.

[119] A. F. Citalán-Madrid et al. Small gtpases of the ras superfamily regulate intestinal epithelial homeostasis and barrier function via common and unique mechanisms. *Tissue Barriers*, 1(5), 2013.

[120] T. A. Martin and W. G. Jiang. Loss of tight junction barrier function and its role in cancer metastasis. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1788(4):872–891, 2009.

[121] B. H. Choi et al. Identification of radil as a ras binding partner and putative activator. *Journal of Biological Chemistry*, 296:100314, 2021.

[122] S. P. Zimmerman et al. Sorting nexin 27 (snx27) associates with zonula occludens-2 (zo-2) and modulates the epithelial tight junction. *Biochemical Journal*, 455(1):95–106, 2013.

[123] X. H. Zhou et al.
Rassf5 inhibits growth and invasion and induces apoptosis in osteosarcoma cells through activation of mst1/lats1 signaling.
Oncology Reports, 32(4):1505–1512, 2014.

[124] I. R. Vetter and A. Wittinghofer.
The guanine nucleotide-binding switch in three dimensions.
Science, 294(5545):1299–1304, 2001.

[125] Y. Zhang and J. Skolnick. Tm-align: a protein structure alignment algorithm based on the tm-score.
Nucleic Acids Research, 33(7):2302–2309, 2005.

[126] J. Mistry et al. Pfam: the protein families database in 2021.
Nucleic Acids Research, 49:D412–D419, D1, 2021.

[127] A. Bateman et al. Uniprot: the universal protein knowledgebase in 2021.
Nucleic Acids Research, 49:D480–D489, D1, 2021.

[128] M. Mirdita et al. Colabfold: making protein folding accessible to all.
Nature Methods, 19(6):679–682, 2022.

[129] R. Evans et al. Protein complex prediction with alphafold-multimer.
bioRxiv:2021.10.04.463034, 2022.

[130] M. Mirdita, M. Steinegger, and J. Söding. Mmseqs2 desktop and local web server app for fast, interactive sequence searches.
Bioinformatics, 35(16):2856–2858, 2019.

[131] M. Mirdita et al. Uniclust databases of clustered and deeply annotated protein sequences and alignments.
Nucleic Acids Research, 45:D170–D176, D1, 2017.

[132] A. L. Mitchell et al. Mgnify: the microbiome analysis resource in 2020.
Nucleic Acids Research, 48:D570–D578, D1, 2020.

[133] H. Berman, K. Henrick, and H. Nakamura.
Announcing the worldwide protein data bank.
Nature Structural & Molecular Biology, 10(12):980–980, 2003.

[134] P. Eastman et al. Openmm 7: rapid development of high performance algorithms for molecular dynamics.
PLOS Computational Biology, 13(7):e1005659, 2017.

[135] M. Ankerst et al.
Optics: ordering points to identify the clustering structure.
ACM SIGMOD Record, 28(2):49–60, 1999.

[136] S. R. Adariani et al. A comprehensive analysis of ras-effector interactions reveals interaction hotspots and new binding partners.
Journal of Biological Chemistry, 296:100626, 2021.

[137] M. W. Dorrity et al. Dimensionality reduction by umap to visualize physical and genetic interactions.
Nature Communications, 11(1):1–6, 2020.

[138] D. F. Burke et al.
Towards a structurally resolved human protein interaction network.
Nature Structural & Molecular Biology 2023 30:2, 30:216–225, 2, 2023.

[139] O. Ursu et al. Massively parallel phenotyping of coding variants in cancer with perturb-seq. *Nature Biotechnology*, 40:896–905, 6, 2022.

[140] A. H. van Boxel-Dezaire, M. R. Rani, and G. R. Stark.
Complex modulation of cell type-specific signaling in response to type i interferons. *Immunity*, 25(3):361–372, 2006.

[141] D. E. Hammond et al. Differential reprogramming of isogenic colorectal cancer cells by distinct activating kras mutations.
Journal of Proteome Research, 14(3):1535–1546, 2015.

[142] S. A. Kennedy et al. Extensive rewiring of the egfr network in colorectal cancer cells expressing transforming levels of krasg13d.
Nature Communications, 11(1):1–14, 2020.

[143] J. M. Shields et al. Understanding ras: 'it ain't over 'til it's over'.
Trends in Cell Biology, 10(4):147–154, 2000.

[144] S. Wohlgemuth et al. Recognizing and defining true ras binding domains i: biochemical analysis.
Journal of Molecular Biology, 348(3):741–758, 2005.

[145] C. Kiel et al. Recognizing and defining true ras binding domains ii: in silico prediction based on homology modelling and energy calculations.
Journal of Molecular Biology, 348(3):759–775, 2005.

[146] E. Castellano and J. Downward.
Ras interaction with pi3k: more than just another effector pathway.
Genes & cancer, 2(3):261–74, 2011.

[147] R. L. Siegel et al. Colorectal cancer statistics, 2020.
CA: A Cancer Journal for Clinicians, 70(3):145–164, 2020.

[148] E. R. Fearon and B. Vogelstein.
A genetic model for colorectal tumorigenesis.
Cell, 61(5):759–767, 1990.

[149] J. G. Tate et al. Cosmic: the catalogue of somatic mutations in cancer.
Nucleic Acids Research, 47:D941–D947, D1, 2019.

[150] D. Romano et al.
Protein interaction switches coordinate raf-1 and mst2/hippo signalling.
Nature Cell Biology, 16(7):673–684, 2014.

[151] H. B. Engin et al. Modeling of ras complexes supports roles in cancer for less studied partners. *BMC Biophysics*, 10(1):1–15, 2017.

[152] M. J. Waldner, S. Foersch, and M. F. Neurath.
Interleukin-6 - a key regulator of colorectal cancer development.
International Journal of Biological Sciences, 8(9):1248–1253, 2012.

[153] J. Zeng et al. Clinicopathological significance of overexpression of interleukin-6 in colorectal cancer.
World Journal of Gastroenterology, 23(10):1780–1786, 2017.

[154] B. Ancrile, K. H. Lim, and C. M. Counter.
Oncogenic ras-induced secretion of il6 is required for tumorigenesis.
Genes & Development, 21(14):1714–1719, 2007.

[155] H. Kikuchi et al. Oncogenic kras and braf differentially regulate hypoxia-inducible factor-1α and -2α in colon cancer. *Cancer Research*, 69(21):8499–8506, 2009.

[156] S. Y. Chun et al. Oncogenic kras modulates mitochondrial metabolism in human colon cancer cells by inducing hif-1α and hif-2α target genes. *Molecular Cancer*, 9(1):1–11, 2010.

[157] S. Zhang et al. Stabilization of hypoxia-inducible factor by dmog inhibits development of chronic hypoxia-induced right ventricular remodeling. *Journal of Cardiovascular Pharmacology*, 67(1):68–75, 2016.

[158] H. H. Hsu et al. Prostaglandin e2-induced cox-2 expressions via ep2 and ep4 signaling pathways in human lovo colon cancer cells. *International Journal of Molecular Sciences*, 18(6):1132, 2017.

[159] N. Smakman et al. Dual effect of krasd12 knockdown on tumorigenesis: increased immune-mediated tumor clearance and abrogation of tumor malignancy. *Oncogene*, 24(56):8338–8342, 2005.

[160] A. Greenhough et al. The cox-2/pge 2 pathway: key roles in the hallmarks of cancer and adaptation to the tumour microenvironment. *Carcinogenesis*, 30(3):377–386, 2009.

[161] M. Y. Hein et al. A human interactome in three quantitative dimensions organized by stoichiometries and abundances. *Cell*, 163(3):712–723, 2015.

[162] A. L. Richards, M. Eckhardt, and N. J. Krogan. Mass spectrometry-based protein–protein interaction networks for the study of human diseases. *Molecular Systems Biology*, 17(1):e8792, 2021.

[163] V. Beltran-Sastre et al. Tuneable endogenous mammalian target complementation via multiplexed plasmid-based recombineering. *Scientific Reports*, 5(1):1–10, 2015.

[164] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674, 2011.

[165] E. L. Huttlin et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, 184(11):3022–3040.e28, 2021.

[166] S. Eisenberg and Y. I. Henis. Interactions of ras proteins with the plasma membrane and their roles in signaling. *Cellular Signalling*, 20(1):31–39, 2008.

[167] K. Miller-Jensen et al. Common effector processing mediates cell-specific responses to stimuli. *Nature*, 448(7153):604–608, 2007.

[168] M. Uhlén et al. Tissue-based map of the human proteome. *Science*, 347(6220), 2015.

[169] E. L. Huttlin et al. The bioplex network: a systematic exploration of the human interactome. *Cell*, 162(2):425–440, 2015.

[170] E. L. Huttlin et al. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017.

[171] M. H. Schaefer and L. Serrano. Cell type-specific properties and environment shape tissue specificity of cancer genes. *Scientific Reports*, 6(1):1–14, 2016.

[172] C. Nogales et al. Network pharmacology: curing causal mechanisms instead of treating symptoms. *Trends in Pharmacological Sciences*, 43(2):136–150, 2022.

[173] C. S. Hughes et al. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, 14(1):68–85, 2018.

[174] J. Rappsilber, M. Mann, and Y. Ishihama. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nature Protocols*, 2(8):1896–1906, 2007.

[175] N. Bache et al. A novel lc system embeds analytes in pre-formed gradients for rapid, ultra-robust proteomics.
Molecular and Cellular Proteomics, 17(11):2284–2296, 2018.

[176] F. Meier et al. Online parallel accumulation–serial fragmentation (pasef) with a novel trapped ion mobility mass spectrometer.
Molecular and Cellular Proteomics, 17(12):2534–2545, 2018.

[177] Y. Perez-Riverol et al. The pride database resources in 2022: a hub for mass spectrometry-based proteomics evidences.
Nucleic Acids Research, 50:D543–D552, D1, 2022.

[178] C. Lazar et al.
Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies.
Journal of Proteome Research, 15(4):1116–1125, 2016.

[179] L. Gatto and K. S. Lilley.
Msnbase-an r/bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation.
Bioinformatics, 28(2):288–289, 2012.

[180] L. Gatto, S. Gibb, and J. Rainer. Msnbase, efficient and elegant r-based processing and visualization of raw mass spectrometry data.
Journal of Proteome Research, 20(1):1063–1069, 2021.

[181] J. Rainer et al. A modular and expandable ecosystem for metabolomics data annotation in r. *Metabolites*, 12(2):173, 2022.

[182] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing.
Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995.

[183] P. J. Luthert and C. Kiel. Combining gene–disease associations with single-cell gene expression data provides anatomy-specific subnetworks in age-related macular degeneration.
Network and Systems Medicine, 3(1):105–121, 2020.

[184] H. Wickham et al. *dplyr: A Grammar of Data Manipulation*. R package version 1.0.9. 2022.

[185] H. Wickham and M. Girlich. *tidyr: Tidy Messy Data*. R package version 1.2.0. 2022.

[186] H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019.

[187] D. Garmonsway. *tidyxl: Read Untidy Excel Files*. R package version 1.0.7. 2020.

[188] L. Henry and H. Wickham. *purrr: Functional Programming Tools*. R package version 0.3.4. 2020.

[189] S. Oba et al. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[190] W. Stacklies et al. Pcamethods—a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23(9):1164–1167, 2007.

[191] C. Lazar and T. Burger. *imputeLCMD: A Collection of Methods for Left-Censored Missing Data Imputation*. R package version 2.1. 2022.

[192] G. Yu et al. Clusterprofiler: an r package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 16(5):284–287, 2012.

[193] T. Wu et al. Clusterprofiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation(China)*, 2(3), 2021.

[194] M. Ashburner et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[195] S. Carbon et al. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49:D325–D334, D1, 2021.

[196] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[197] A. Schlicker et al. A new measure for functional similarity of gene products based on gene ontology.
BMC Bioinformatics, 7(1):1–16, 2006.

[198] G. Yu et al. Gosemsim: an r package for measuring semantic similarity among go terms and gene products.
Bioinformatics, 26(7):976–978, 2010.

[199] Z. Gu and D. Hübschmann. Simplify enrichment: a bioconductor package for clustering and visualizing functional enrichment results.
Genomics, Proteomics & Bioinformatics, 2022.

[200] Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data.
Bioinformatics, 32(18):2847–2849, 2016.

[201] W. Chang et al. *shiny: Web Application Framework for R*.
R package version 1.7.1. 2021.

[202] D. Attali. *shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*. R package version 2.1.0. 2021.

[203] W. Chang and B. Borges Ribeiro.
shinydashboard: Create Dashboards with 'Shiny'.
R package version 0.7.2. 2021.

[204] Z. Gu and D. Hübschmann. Make interactive complex heatmaps in r.
Bioinformatics, 38(5):1460–1462, 2022.

[205] C. R. Harris et al. Array programming with {numpy}.
Nature, 585:357–362, 2020.

[206] P. Virtanen et al.
{Scipy} 1.0: fundamental algorithms for scientific computing in python.
Nature Methods, 17:261–272, 2020.

[207] The pandas development team. Pandas-dev/pandas: pandas, 2020.

[208] W. McKinney et al. Data structures for statistical computing in python.
In volume 445, pages 51–56, 2010.

[209] S. K. Lam, A. Pitrou, and S. Seibert.
Numba: a llvm-based python jit compiler. In pages 1–6, 2015.

[210] D. Vaughan and M. Dancho.
furrr: Apply Mapping Functions in Parallel using Futures.
R package version 0.3.0. 2022.

[211] T. L. Pedersen. *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks*. R package version 2.0.5. 2021.

[212] T. Will and V. Helms. Rewiring of the inferred protein interactome during blood development studied with the tool ppicompare.
BMC Systems Biology 2017 11:1, 11:1–19, 1, 2017.

[213] Z. Liu et al.
A large accessory protein interactome is rewired across environments.
eLife, 9:1–65, 2020.

[214] C. Ternet. Conditions-specific quantitative network rewiring of colon cancer-associated kras mutations in caco-2 cell line, 2022.

[215] S. H. Lee, M. Golinska, and J. R. Griffiths. Hif-1-independent mechanisms regulating metabolic adaptation in hypoxic cancer cells.
Cells 2021, Vol. 10, Page 2371, 10:2371, 9, 2021.

[216] R. Chen, M. A. Ahmed, and N. R. Forsyth.
Dimethyloxalylglycine (dmog), a hypoxia mimetic agent, does not replicate a rat pheochromocytoma (pc12) cell biological response to reduced oxygen culture.
Biomolecules 2022, Vol. 12, Page 541, 12:541, 4, 2022.

[217] E. P. Cummins et al. The hydroxylase inhibitor dimethyloxalylglycine is protective in a murine model of colitis.
Gastroenterology, 134:156–165.e1, 1, 2008.