Regulation Flow Analysis discovers molecular mechanisms of action from large knowledge databases

Carlos P. Roca[1], Oleg Sysoev[2], Elena Eyre[3], Silvia Galan[3], Dominic Sinibaldi[4], Philip Tedder[1], and Jonathan Mangion[1]

[1] *DS&AI, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK*
[2] *Department of Computer and Information Science, Linköping University, Linköping, Sweden.*
[3] *DS&AI, BioPharmaceuticals R&D, AstraZeneca, Barcelona, Spain*
[4] *DS&AI, BioPharmaceuticals R&D, AstraZeneca, Gaithersburg, US*

Abstract

Drug development is a long and expensive process, with only a small fraction of potential drugs being finally approved. The challenge of drug development is rooted in our limited understanding of biological systems and the disease processes that drugs are trying to modulate. We propose a novel method, called Regulation Flow Analysis (RFA), which is based on the principles of biological regulation, causal graphs, and graph flow. RFA is able to generate causal hypotheses of mechanism of action, using large Knowledge Graphs (KG) of molecular regulation. Our numerical experiments demonstrate that the generated hypotheses, in the form of regulation pathways, often summarize mechanisms of drug action in a manner both understandable and actionable. Thus, RFA can greatly improve our understanding of the biological processes underlying health and disease, and therefore substantially facilitate drug development. Our examples illustrate how RFA recovers known mechanisms of action and can be used for target and biomarker discovery and validation.

Keywords: regulation, causal, graph, network, mechanism of action, pathomechanism, genetic variant, drug, target, biomarker

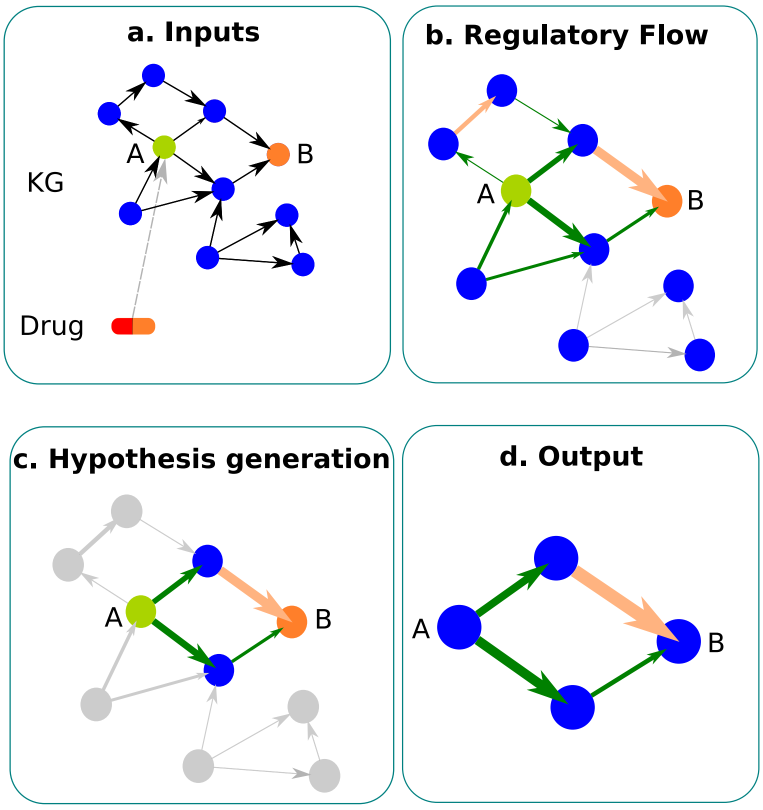

Figure 1: Graphical abstract of the paper: **a.** The Knowledge Graph (KG), the target of the drug (A), and the endpoint of the pathway (B) are provided as inputs; **b.** Regulation Flow is computed in the KG; **c.** Mechanistic hypothesis generation (Algorithm 3) computes the most informative pathways between A and B; **d.** The resulting pathways are output for interpretation and decision making.

1 Introduction

The process of drug discovery and development is both long and expensive, with an average time of 15 years for a drug to reach market, at a cost of approximately two billion dollars [1]. Less than 10% of new drugs in Phase 1 will eventually be launched, with the majority of failures happening in the transition from Phase 2 to Phase 3, because of toxicity or lack of efficacy [2, 3]. The difficulty of drug development is linked to the inherent complexity and incomplete understanding of the biological systems that are being targeted. If reliable hypotheses of the mechanisms of action of potential drug targets can be developed and tested, this will lead to higher success rates due to better target discovery and validation. Additionally, it would lead to the selection of more accurate biomarkers and the better prediction of possible side effects.

AstraZeneca applies the 5R framework to R&D [4]: right target, right tissue, right safety, right patient, and right commercial potential. Target selection remains the most important decision in the drug discovery process, and interrogating the growing knowledge of disease biology is critical to making the right choice. Information and knowledge about diseases, targets and drugs is available from many different sources, but integrating this biomedical knowledge is complex and challenging. Here we focus on the critical aspect of improving our understanding of the underlying disease processes and the ways to modulate them, by generating accurate pathways between genes or proteins, as causal graphs representing biological regulation. This allows the rational selection of drug targets, the prediction of downstream events of modulating those targets, and the identification of potential biomarkers.

Regulation networks can be represented by directional and causal graphs, which provide the proper conceptual framework to model known interactions between biological entities, such as protein-protein interactions, post translational modifications, and transcriptional regulation. These forms of knowledge graphs (KGs) can be readily built from available data resources, both academic (for example, OmniPath [5] or Reactome [6]) and commercial (such as Clarivate MetaBase [7] , Qiagen Knowledge Base [8], or Metaphacts Knowledge Graph [9]). These knowledge bases collect large amounts of information in the literature and in curated databases about biological regulation at molecular level. Being very information rich, such large KGs have a great potential for revealing the mechanisms of action of potential drug targets.

Signaling in large KGs has been modeled by a number of approaches that integrate omics data [10, 11, 12, 13, 14, 15, 16]. Ingenuity Pathway Analysis (IPA) [17] computes upstream regulators of the given gene set and generates mechanistic hypotheses connecting the upstream regulators with this gene set. Similar principles are also used for identification of the downstream targets in IPA. When it comes to the methods that generate mechanistic hypotheses purely from KGs, the literature is very scarce. Nichenet applies Dijkstra's shortest path algorithm to infer signaling paths between a ligand and the target gene of interest [14]. Signal Flow Control (SFC) algorithm [18, 19] introduces a dynamic model with a combination of network signal propagation and basal activation, and it uses the gradients of the network flow to estimate the signal in the steady state. This method demonstrates that the sign of the predicted effect of a given perturbation agrees with the direction of the experimentally observed differential expression changes in 60-80% of genes. While the objective of SFC is not the generation of mechanistic hypotheses, it illustrates that algorithms on regulation graphs are able to explain the effects observed experimentally to a reasonable extent.

In this paper, we present a novel method, called Regulation Flow Analysis (RFA), developed and actively used in AstraZeneca for generating hypotheses of mechanism of action at molecular level, by identifying the causal events connecting sources of perturbation (i.e. a deleterious genetic variant, or a ligand-receptor binding) to the observed effects (a change in gene/protein expression, or a change in a molecular biomarker). Our model consists of two parts: the Regulation Flow model and the algorithm for generating hypotheses of mechanism of action. While previous work has explored the potential of graph flow to model biological regulation [18, 19], we consider the algorithm to generate mechanistic hypotheses as the main novel contribution reported in this paper. We apply our method to predicting the underlying biology of Noonan syndrome, thus demonstrating the capacity of RFA for generating mechanistic hypotheses, and to identifying drug targets for a developmental disease, Fragile X syndrome.

2 Regulation Flow model

To model regulation in KGs, we propose a model which we call Regulation Flow model. The model assumes that after one or a few vertices in the KG are perturbed, a cascade of regulation effects propagates throughout the graph. The model does not have a temporal meaning, but rather it represents the steady state some time after the perturbation happened. The effect on any given vertex will be, by assumption, the addition of the effects arriving by all the possible different paths in the graph connecting the initially perturbed vertices to that particular vertex.

In mathematical terms, we have a regulation graph $G =< V, E >$, where $V = \{v_1, \ldots, v_n\}$ is the set of graph vertices or nodes, and $E = \{e_{ij} = (v_i, v_j); v_i, v_j \in V\}$, is the set of graph edges or links, with associated labels l_{ij}. In general, vertices represent genes or proteins, whereas edges represent direct regulation relationships, at molecular level, between them. When $l_{ij} = 1$, we have positive regulation, this meaning that up-regulation of v_i leads to the up-regulation of v_j, and conversely down-regulation of v_i leads to the down-regulation of v_j. On the other hand, $l_{ij} = -1$ indicates negative regulation, so that if v_i is up-regulated, then v_j is down-regulated, and conversely down-regulation of v_i causes up-regulation of v_j. Up- and down-regulation have here the usual meaning in biological systems, that is, stimulation and inhibition, respectively.

Given these inputs, we compute a signed adjacency matrix $A = (A_{ij})$ from the graph G, following the standard definition, $A_{ij} = l_{ij}$ if $(v_i, v_j) \in E$ and $A_{ij} = 0$ otherwise. By using the adjacency matrix, we compute an *unscaled normalized step matrix* $U = (U_{ij})$ as the transpose of the A matrix, normalized by the number of outgoing edges of each vertex:

$$U_{ji} = \frac{A_{ij}}{\sum_{k=1}^{n} |A_{ik}|}. \tag{1}$$

The rationale of this normalization is that vertices (genes or proteins) with effects on many others vertices, will tend to have effects of smaller magnitude or strength, in comparison to vertices affecting a lower number of vertices.

This unscaled step matrix represents the propagation of a regulation effect in a single step, or in other words, over a path of length one in the graph. In an experiment or intervention, we are given a perturbation set with one or a few vertices, $\{v_{i_1}, \ldots, v_{i_k}\}$, describing the vertices that are the origin or source of the perturbation. We codify this perturbation with the vector $p = (p_t)$, with $p_t \neq 0$ when $t \in \{i_1, \ldots, i_k\}$ and usual values $|p_t| = 1$. For example, a single perturbation may be a drug which stimulates a receptor, which in another intervention may be combined with a compound that inhibits a kinase. Let $x = (x_1, \ldots, x_n)$ represent the resulting regulation effect, so that $x_i > 0$ if the perturbation p up-regulates vertex v_i, $x_i < 0$ if it down-regulates vertex v_i, and $x_i = 0$ otherwise. Then, the one-step propagation of regulation can be calculated as:

$$x^{(1)} = U \cdot p. \tag{2}$$

Let us also define $x^{(0)} = p$, the perturbation itself, without effects on any other vertex.

The regulation reaching a vertex after traversing m steps in the graph can also be calculated as:

$$x^{(m)} = U^m \cdot p. \tag{3}$$

The *total regulation* on any vertex, composed by the regulation reaching via paths of any length, is accordingly:

$$x = x^{(0)} + x^{(1)} + x^{(2)} + \ldots = p + U \cdot p + U^2 \cdot p + \ldots = (I + U + U^2 + \ldots) \cdot p, \tag{4}$$

I being the identity matrix.

Since (4) includes the sum of an infinite series, convergence criteria must be verified. The sum of the matrix powers $I + A + A^2 + \ldots$ is a Neumann series ([20], page 618), whose necessary and sufficient criterion of convergence is that the spectral radius $\rho(U)$ (the modulus of the largest eigenvalue) is strictly less than one. This can not be guaranteed for matrix U in general, and therefore we introduce the *scaled step matrix* S as follows:

$$S = \frac{\alpha}{\rho(U)} U, \tag{5}$$

where $0 < \alpha < 1$, for example, $\alpha = 0.99$. As by definition $\rho(S) = \alpha$, this guarantees $\rho(S) < 1$. Accordingly, we modify equation (4) as follows:

$$x^* = (I + S + S^2 + \ldots) \cdot p = (I - S)^{-1} p. \tag{6}$$

This way, the parameter α represents the scale of the one-step effect produced by perturbing with the largest eigenvector of the step matrices U or S. Therefore, it controls the overall scale of the response to perturbations in the model.

The last part of equation (6), as well as the existence of the inverse, follows from the properties of the Neumann series ([20], page 618). We call the matrix F the *regulation flow matrix*,

$$F = (I - S)^{-1}, \tag{7}$$

and x, the regulation effect caused by perturbation p, corresponds to the *total regulation flow*,

$$x^* = Fp. \tag{8}$$

Each coefficient of the regulation flow matrix, F_{BA}, represents the regulation flow from A to B, that is, the total regulation influence of gene/protein A over gene/protein B. A compact description of the Regulation Flow model is provided in Algorithm 1.

Algorithm 1 Regulation flow model

```
Input: Prior Knowledge graph $G$, perturbation $p$
Compute adjacency matrix $A$ from $G$
Compute unscaled step matrix $U$ by (1)
Compute scaled step matrix $S$ by (5)
Compute flow matrix $F$ by (7)
Predict the flow $x$ with (8)
Output: $x^*$
```

3 Generation of mechanistic hypotheses by RFA

Pathways are essential constructs in biology and biomedicine, conceived to organize biological processes as series of events describing the causal links between an initial perturbation and a series of downstream effects. By perturbation we mean, for example, the activation of a kinase, the coupling of a ligand with its receptor, or the formation of a protein complex. Here, we formalize the concept of a pathway with a precise definition. As a first step, we will define the *pathway from gene/protein A to gene/protein B* as the subgraph defined by all the regulation paths connecting A to B. In most practical applications such a pathway would contain a very large number of intermediate genes/proteins and regulation paths, the latter frequently infinite because of the presence of cycles. However, as we will show, a crucial property is that in most cases only a very small number of intermediate genes/proteins and regulation paths channel almost all of the regulation flow from A to B. This allows to consider the much simpler *effective pathway* between A and B, instead of the full pathway.

Let us start considering the intermediate genes/proteins $\{C_1, C_2, \ldots, C_n\}$ between the genes/proteins A and B. For any intermediate gene/protein C_i, there exists at least one path connecting A to C_i, and at least one path connecting C_i to B. Any path P from A to B traversing through C_i will be defined by the sequence of vertices $P = (A, D_1, D_2, \ldots, D_n, B)$, with at least one $j = 1 \ldots n$ such that $D_j = C_i$. We call this path C_i-acyclic if C_i occurs only once in the sequence P. In general, let us consider k_1 to be the index of the first occurrence of vertex C_i in P, and let k_2 be the index of the last occurrence of C_i in P. If $k_1 = k_2$, then P is C_i-acyclic. Otherwise, when $k_1 \neq k_2$, P is a concatenation of $P_1 = (A, D_1, \ldots, D_{k_1})$, $P_2 = (D_{k_1}, \ldots, D_{k_2})$ and $P_3 = (D_{k_2}, \ldots, B)$, where P_1 and P_3 are C_i-acyclic and P_2 is a cycle from C_i to C_i, which we call a C_i-cycle. C_i-cycles may have any finite number of visits to vertex C_i.

From this, it is clear that all the paths from A to B that traverse through C_i can be obtained as the composition of all the C_i-acyclic paths from A to C_i, all the C_i-cycles, and all the C_i-acyclic paths from C_i to B. By making $C_i = A$ or $C_i = B$, we can also state that all the paths from A to B can be obtained as the composition of all the A-cycles and all the A-acyclic paths from A to B, or as the composition of all the B-acyclic paths from A to B and all the B-cycles.

Algebraically, the total regulation flow from A to B is equal to the sum of the flow contributed by each of the connecting paths (4). Additionally, the flow channeled by any path P can be factored into the multiplication of the flow channeled by any two subpaths P_1 and P_2 such that P is the concatenation of P_1 and P_2. As a result, by the distributive property, the composition of paths corresponds algebraically to the multiplication of flow.

Let us denote by $F_{BA}^{(C_i)}$ the flow from A to B channeled via an intermediate gene/protein C_i. Also, let us denote by $F_{BA}^{(C_i)^*}$ the flow from A to B channeled via C_i, but only through C_i-acyclic paths. From the composition properties above, we can write

$$F_{C_iA} = F_{C_iA}^{(C_i)^*} F_{C_iC_i}, \tag{9}$$

$$F_{BC_i} = F_{C_iC_i} F_{BC_i}^{(C_i)^*}, \tag{10}$$

$$F_{BA}^{(C_i)} = F_{C_iA}^{(C_i)^*} F_{C_iC_i} F_{BC_i}^{(C_i)^*}, \tag{11}$$

which implies

$$F_{BA}^{(C_i)} = \frac{F_{C_iA} F_{BC_i}}{F_{C_iC_i}}. \tag{12}$$

Therefore, the regulation importance $R_{BA}^{(C_i)}$ of any intermediate gene/protein C_i in the pathway from A to B, can be defined as the absolute value of the ratio between the regulation channeled through the intermediate gene/protein and the total regulation,

$$R_{BA}^{(C_i)} = \left| \frac{F_{BA}^{(C_i)}}{F_{BA}} \right| = \left| \frac{F_{C_iA} F_{BC_i}}{F_{BA} F_{C_iC_i}} \right|. \tag{13}$$

This quantity can be easily calculated, as the terms on the r.h.s. can be read directly from the flow matrix F (4).

The relative importance of an intermediate gene/protein allows to design a greedy algorithm to simplify the pathway from A to B to a collection of successively more precise effective pathways, but comparatively much smaller than the full pathway from A to B.

Algorithm 2 Effective pathway between genes/proteins A and B

```
Input: Regulation graph $G =< V, E >$, Flow matrix $F$.
Identify all effective intermediate genes/proteins $\{C_1, C_2, ..., C_n\}$, as those that verify $F_{BA}^{(C_i)} \neq 0$.
Sort intermediate proteins by decreasing regulation importance $R_{BA}^{(C_i)}$, $\{C_{(1)}, C_{(2)}, ..., C_{(n)}\}$ so that $R_{BA}^{(C_{(i)})} \geq R_{BA}^{(C_{(i+1)})}$.
Define the regulation subgraph $G_0$, with only the proteins/genes $A$ and $B$, as the subgraph induced in $G$ by $V_0 = \{A, B\}$.
repeat
    Define the regulation subgraph $G_i$ by adding gene/protein $C_{(i)}$ to $G_{i-1}$, that is, as the subgraph induced in $G$ by $V_i = \{C_{(i)}\} \cup V_{i-1}$.
    Calculate the regulation flow from $A$ to $B$ in subgraph $G_i$, $F_{BA}^{[i]}$, with (7).
    if $F_{BA}^{[i]} \neq F_{BA}^{[i-1]}$ then
        Output subgraph $G_i$ and flow $F_{BA}^{[i]}$.
    end if
until $F_{BA}^{[i]}$ close enough to $F_{BA}$ or $i$ is large enough.
```

As the edge weights in G are signed, some paths may contribute flow which opposes (opposite sign) the overall flow. In those cases the increase in the finite sequence $(F_{BA}^{[i]})$ is not monotonous, but convergence is ensured as by construction $F_{BA}^{[n]} = F_{BA}$. In the initial steps $i = 0, 1, 2, ...$, the subgraph G_i may be

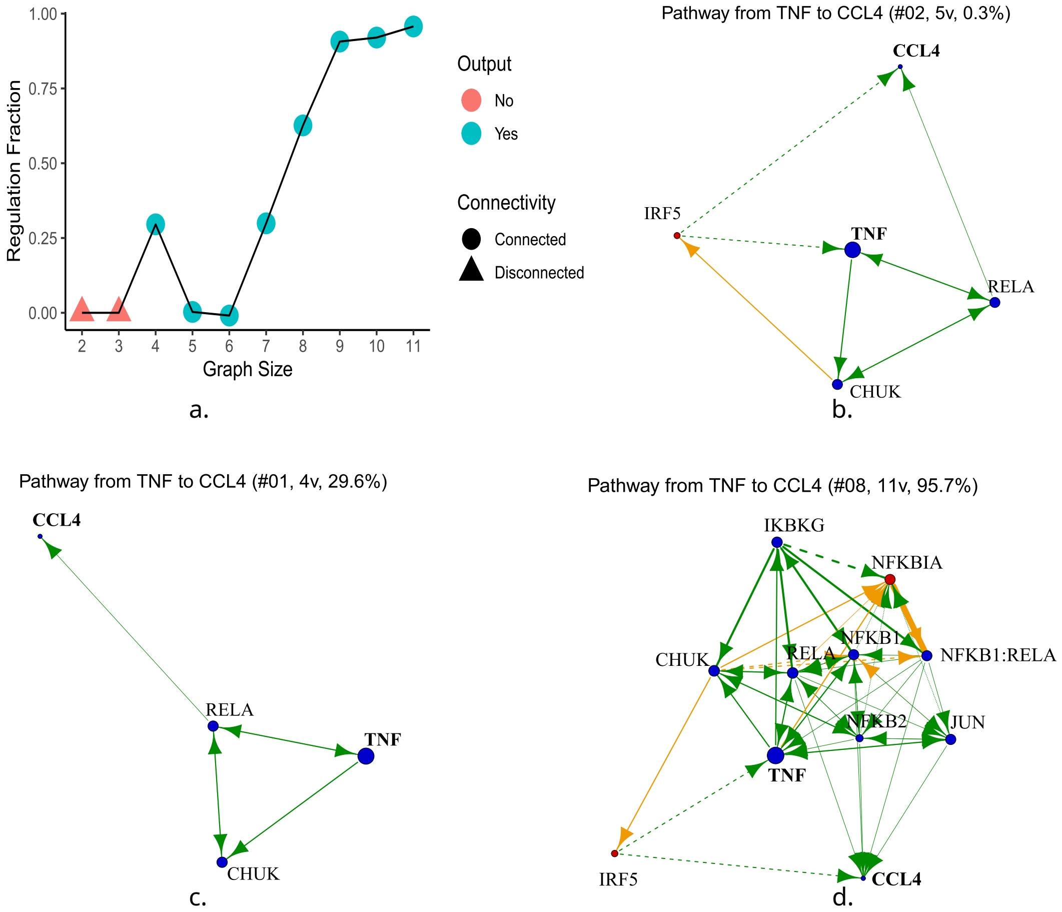

Figure 2: **a.** Dependence of the Regulation Fraction on the subgraph size. When a subgraph is disconnected, it is not output in our implementation. **b.-d.** Pathways between TNF and CCL4 with 4, 5 and 11 vertices, respectively. The title is reporting the iteration number (#XX), number of vertices (XXv) and regulation percentage of the given subgraph. Vertices represent genes/proteins, and edges represent direct regulation between genes/proteins. Vertex sizes display the amount of total regulation caused by the original perturbation, TNF. Vertex colors indicate if vertices are regulated in the same way as TNF (blue) or in the opposite way (red), that is, up- or down-regulated, respectively. Edge widths represent the amount or strength of regulation. Edge colors indicate the sign of regulation, positive (green) or negative (orange).

disconnected, and thus $F_{BA}^{[i]} = 0$. The subgraphs G_i may also contain *loose branches*, that is, series of interconnected genes/proteins D_j not connected in the subgraph to A or B. Those loose branches are easily identified, as $F_{D_jA}^{[i]}$ or $F_{BD_j}^{[i]}$ will be zero, and they are no pictured in the generated subgraph.

In summary, after defining the pathway from gene/protein A to gene/protein B as the subgraph induced by all the intermediate genes/proteins between A and B, a remarkable capability of graph flow is the calculation of the flow channeled through any intermediate gene/protein. This, in turn, allows to rank intermediate genes/proteins by the amount of regulation channeled through them in the pathway, leading to a greedy algorithm that iteratively builds the much smaller, but functionally equivalent, effective pathways from A to B.

3.1 Illustration of the generation of mechanistic hypotheses

To illustrate the algorithm for generating effective pathways, we show how the regulation $F_{AB}^{[i]}$ in the successive subgraphs G_i approaches F_{AB} (see Algorithm 3), in the case of the pathway between TNF and CCL4, i.e. when $A = \text{TNF}$ and $B = \text{CCL4}$.

Figure 2a shows that the regulation fraction $F_{AB}^{[i]}/F_{AB}$ is equal to zero for the initial subgraph G_0 with only the two vertices A and B, and then it increases non-monotonically to a value close to 1 when the subgraph contains 11 vertices. Thus, this small subgraph explains the regulation flow of the entire subgraph between TNF and CCL4, which has 3,667 vertices, with an accuracy larger than 95%. Figures 2b-2d picture the actual subgraphs constructed by the algorithm at various iterations.

We have observed in our numerical experiments that the regulation fraction changes non-monotonically, and sometimes the regulation fraction can exceed 1. This happens due to the signed nature (stimulation or inhibition) of the biological regulation modeled by the graph: when edges with opposite sign are added in subsequent iterations of the algorithm, newly added paths may channel flow of opposite sign, which counteracts the regulation flow between A and B so far. This can be observed for example with the subgraphs #01 and #02 in Figs. 2b and 2c.

4 Results

Validation of RFA is shown by demonstrating the mechanism of action of genes that have been implicated in causing Noonan Syndrome and related disorders. Syndromes including Noonan Syndrome, Costello syn-

Table 1: Noonan-like syndromes and the genes that have been implicated in the disease

RASopathy	**Genes**	**Mutation Effect**	**Pred. reg. on MAPK1**	**Reference**
Noonan Syndrome	*PTPN11*	Gain of function	-	[22]
	KRAS	Gain of function	+	[23]
	SOS1	Gain of function	+	[24]
	RAF1	Gain of function	+	[25]
	NRAS	Gain of function	+	[26]
	BRAF	Gain of function	+	[27]
	RIT1	Gain of function	+	[28]
	SOS2	Gain of function	+	[29]
	LZTR1	Loss of function	-	[30]
	MRAS	Gain of function	+	[31]
	RRAS	Gain of function	+	[32]
	MAPK1	Gain of function	+	[33]
	SPRED2	Loss of function	Not available	[34]
Noonan Syndrome - Neurofibromatosis Type 1	NF1	Loss of function	-	[35]
LEOPARD Syndrome	*PTPN11*	Dominant negative	-	[36]
	RAF1	Gain of function	+	[25]
	BRAF	Gain of function	+	[27]
Costello Syndrome	*HRAS*	Gain of function	+	[28]
Cardiofaciocutaneous syndrome	*BRAF*	Gain of function	+	[27]
	KRAS	Unconfirmed	+	[37]
	MAPK1	Gain of function	+	[38]
	MAPK2	Gain of function	+	[38]
Legius Syndrome	*SPRED1*	Loss of function	Not available	[39]
Noonan syndrome-like disorder with loose anagen hair	*SHOC2*	Loss of function	+	[40]
	PPP1CB	Unconfirmed	-	[41]
Noonan syndrome-like disorder with or without juvenile myelomonocytic leukaemia	*CBL*	Loss of function	-	[42]
Noonan syndrome-like disorder with or without craniosynostosis	*ERF*	Loss of function	-	[43]

drome, LEOPARD syndrome and neurofibromatosis type 1 are caused by germline mutations in genes that cause dysregulation of the RAS-MAPK pathway and are collectively know as RASopathies [21]. In addition, an example of applying RFA in a pharmaceutical setting is provided by exploring targets for Fragile X syndrome.

4.1 RFA recovers known mechanisms of action of RASopathies

Signaling through the RAS-MAPK pathway plays a critical role in cell differentiation, proliferation and survival. The various mutations in 22 genes that have been associated with RASopathies (Table 1) cause an up-regulation of the RAS-MAP pathway, with *MAPK1* and *MAPK2* being dysregulated downstream. Here we use the known biology of RASopathies to validate RFA, by confirming that the method can reconstitute the expected pathway and predict how the mutations act to cause overall downstream dysregulation of *MAPK1*.

The results from RFA are shown in Figure 3. No pathways between *SPRED1* or *SPRED2* to *MAPK1* were generated, as they were not annotated in OmniPath. The results confirm the expected mode of regulation of *MAPK1*, except in the case of two genes; *PTPN11* and *SHOC2*. RFA predicts that overall *PTPN11* has a negative effect on the regulation of MAPK1. The *PTPN11* mutations that cause Noonan syndrome have been shown to be gain of function mutations, which overstimulate the MAPK/ERK pathway. Kontaradis et al. [36] demonstrate that in LEOPARD syndrome the mutations are limited to the PTP domain and have a dominant negative effect by inactivating the catalytic function. This raises the question of how gain of function and dominant negative mutants can result in a similar phenotype. Kontaradis et al. argue that Noonan syndrome and LEOPARD syndrome phenotypes result from differential effects of mutant Shp2 on different receptor-tyrosine kinase pathways at distinct developmental times. The pathway derived by RFA (Figure 3b) indicates two possible routes of regulation between *PTPN11* and *MAPK1.*, one of which acts through HRAS and would require a gain of function mutation in *PTPN11*, while the second pathway requires *VAV1*. AS *PTPN11* inhibits the action of *VAV1*, which in turn up-regulates *MAPK1*, a loss of function within *PTPN11* would result in increased activity of both *VAV1* and *MAPK1*. It is interesting to hypothesise that in LEOPARD syndrome it is the *PTPN11*/*VAV1* pathway that is regulating *MAPK1*. RFA predicts that overall *SHOC2* up-regulates *MAPK1*, although Cordeddu describe the mutations as loss of function. The shortest pathway generated by RFA illustrates *SHOC2* acting through *PPP1CA* to down-

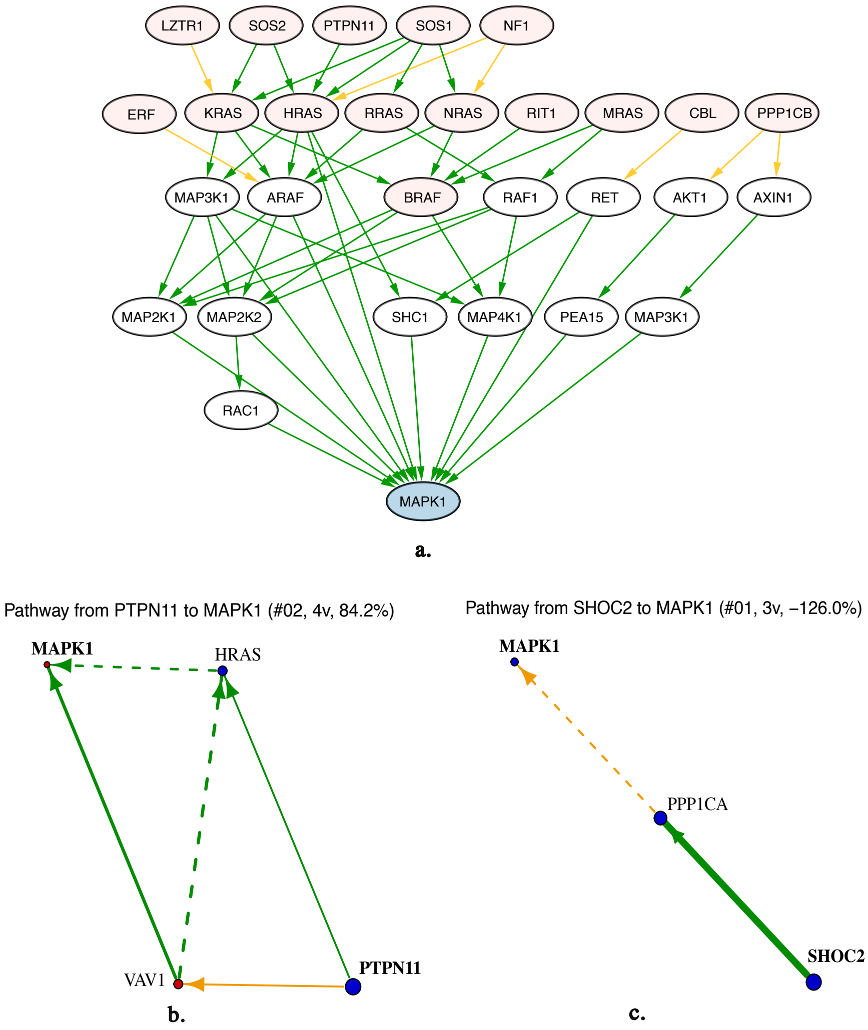

Figure 3: RFA analysis predicts the mutation effect in genes causing RASopathies. **a.** This figure, generated by Cytoscape with RFA results (by integration of the vertices and edges from different RFA results), illustrates interconnected signaling pathways involved in RASopathies, and highlights critical components, including the RAS family to MAPK pathway (pink and blue vertices). Dysregulation of these pathways leads to the pathogenesis of RASopathies, such as Noonan syndrome. Pink vertices in upstream position correspond to genes with known association to RASopathies. We used those genes as source genes in RFA, and MAPK1 as the target gene. Green lines indicate positive regulation, yellow lines indicate negative regulation. Arrows indicate flow direction. **b.** and **c.** show the pathways produced by RFA from PTPN11 and SHOC2, respectively (sizes, widths, and colors of vertices and edges are as defined in Fig. 2)

.

regulate *MAPK1* (Figure 3c). In this case, the results agree with the literature. The graph shows a dotted yellow line between *PPP1CA* and *MAPK1*, as the predicted inhibition is opposite to the overall effect. Adding extra nodes to the graph shows that there are alternative pathways between *SHOC2* and *MAPK1*, and we hypothesize that the pathway through *PPP1CA* shown in Fig. 3c is the one causing the RASopathy.

Functional studies have not been performed on the mutations in *PPP1CB*, whereas RFA results predict that loss-of-function mutations would result in up-regulation of ERK. The recent discovery that loss-of-function mutations in *ERF* cause a Noonan-syndrome-like disorder, with or without craniosynostosis, led Dentici et al. [43] to suggest that this disorder should be classified as a RASopathy. The results shown in Fig. 3a confirm that the mode of action is similar to other RASopathies and support the suggestion by [43].

4.2 RFA discovers targets for Fragile X syndrome

Fragile X syndrome (FXS) is caused by mutations in the *FMR1* gene, leading to a non-functional FMRP protein. Gene therapy has been proposed as a way to recover the expression of FMRP. From a drug development and commercial perspective, gene therapy is expensive, and in the case of FXS the therapy would need to be delivered across the blood-brain barrier. We used RFA to investigate whether there were other possible drug target candidates within the regulatory pathway downstream of *FMR1*. Brain-derived neurotrophic factor (BDNF) has been proposed as a biomarker for FXS [44], and it has been shown to be dysregulated in *FMR1* knock-out mice [45]. Therefore we investigated how *FMR1* regulates *BDNF* expression, with the aim of uncovering as potential targets the main intermediate genes/proteins channeling this regulation.

Figure 4 shows that *FMR1* up-regulates *BDNF* expression through *PAK1* and *DLG1*. *DLG1* encodes for a scaffold protein that contains a PDZ domain and is implicated in synaptic activity. Small molecule and peptide inhibitors of PDZ activity have been developed for neurological and cancer indications [46]. *PAK1*

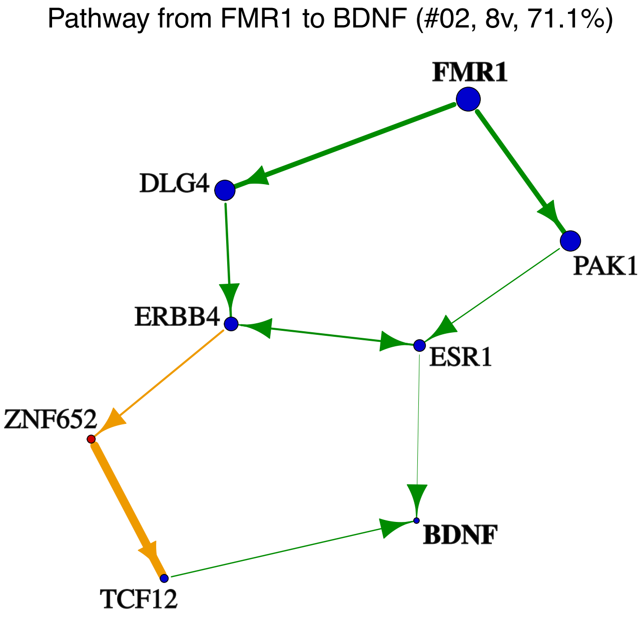

Figure 4: Regulation pathway from *FMR1* to *BDNF*. Sizes, widths, and colors of vertices and edges are as defined in Fig. 2.

encodes for a protein kinase known to play a wide role in cell signaling through its catalytic and scaffolding activities, and it is known to be dysregulated in various neurological disorders including FXS [47]. PAK-inhibiting compounds have been developed for both neurological disorders and cancer [48]. Genentech has developed PAK-inhibitors specifically to treat FXS, which permeate across the blood-brain barrier and ameliorate the FXS phenotype in *Fmr1* knock-out mice [49].

In summary, the pathway generated by RFA shows how *FMR1* regulates *BDNF* and allows for the identification of intermediate proteins that have the potential to act as drug targets. Importantly, the causal and molecular nature of the hypotheses generated by RFA allows for clearly defined experimental validation.

5 Discussion

We have presented here the Regulation Flow Analysis (RFA) model, which aims to quantify the amount of regulation between two genes or proteins, by making use of regulation graphs and their flow properties. By using the RFA model, we have devised a greedy algorithm for generation of mechanistic hypotheses. One of the main findings of our paper is that this algorithm is often able to construct *effective pathways* between two biological units of interest, which are much smaller than the full pathway between them, while representing a similar amount of regulation. For example, the full pathway from TNF to CCL4 includes 3,367 genes/proteins, but more that 95% of regulation in this pathway can be explained by a pathway with only 11 genes/proteins. This demonstrates that our algorithm can be seen as a powerful tool for generating understandable and testable models of mechanisms of action.

The application of graph flow carries an implicit assumption of linearity, in the sense that direct regulation between any two connected genes/proteins can be quantified by the multiplication of a factor (the edge weight), and that the regulation over any gene/protein from several direct upstream connections can be calculated by their sum. Apart from the edge weights, RFA has a main parameter, α in Eq. (5), which ensures a realistic model with finite total regulation. Qualitatively, this parameter reflects the overall scale of the response produced by perturbations.

Our proposed method, RFA, and the Signal Flow Control (SFC) method published previously [18, 19] have certain commonalities and differences. Based on principles of signal propagation and gradient computations, the SFC authors derive Eq. (7) in [19], which seems similar to our Eq. (6). However, the modeling approaches of RFA and SFC are different. On one hand, RFA aims to model biological regulation of any kind, not only signaling processes. On the other hand, RFA models the effect of a given perturbation simply as the propagation of regulation, whereas SFC models this effect as a user-defined combination (sum) of the propagation of regulation and basal activity. As a consequence, SFC has two user-defined hyperparameters (α and β Eq. (7) in [19]) defining that combination, while RFA has none of this. The authors of SFC note that their flow computation is not possible directly sometimes. Indeed, in some cases the user-defined hyperparameters may lead to the scaled step matrix having a spectral radius larger than 1, and therefore its inverse will not exist. Possibly due to this, SFC authors propose an iterative algorithm involving a repeated matrix multiplication (Table 3 of [19]), rather than the direct inverse computation carried out in RFA. Depending on the amount of iterations, this may lead to a much greater computational cost compared to the direct inversion, and it also may lead to numerical overflow if the spectral radius of the scaled step matrix is close to 1. Another, less important, difference between RFA and SFC is the different scaling of the step matrix, using only out-degree (RFA) or both in-degree and out-degree (SFC).

We have performed some comparison of performance with real datasets, using the default hyperparameters values in SFC, which emphasize regulation flow over basal activation. As expected with that setting, results show that the quality of prediction of RFA and SFC, in terms of sign and magnitude of perturbation effects

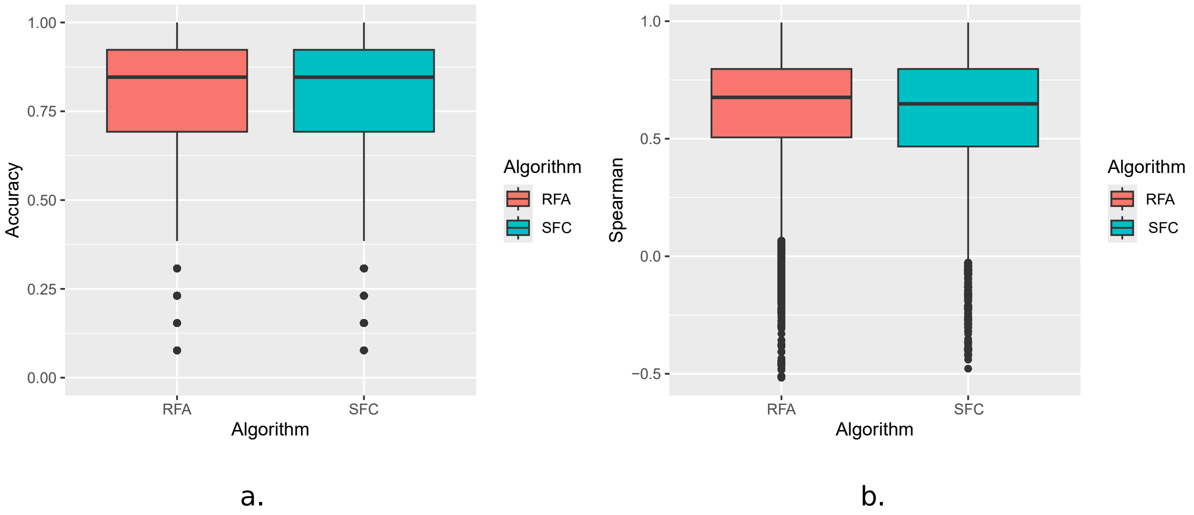

Figure 5: Comparison of RFA and SFC methods in terms of: **a.** accuracies of sign of predicted expression changes, and **b.** Spearman's correlation between observed and predicted expression changes.

over a large set of genes, are similar: one model performed better in some scenarios, while the other model had better performance in others, resulting in similar overall performance of the two methods (see Supplementary Information).

The two provided examples about RASopathies and Fragile X Syndrome showcase the potential of RFA in drug research and development. The molecular nature of regulation graphs, together with their directionality and signed nature, establish a general framework for understanding the causal relationships between genetic variation, disease mechanisms, and biomarkers. As a consequence, it facilitates the rational discovery and validation of disease targets, in the form of genes/proteins predicted to have a therapeutically beneficial regulation influence on the core events driving a disease. Compared to Bayesian graphs, where edge weights represent conditional probabilities, regulation graphs provide a more adequate framework to model biological regulation, given the built-in flexibility in the sign and weight of edges.

RFA, like other models based on regulation graphs, is highly dependent on a good estimation of the edge weights, that is, on the proper estimation of the strength of regulation between genes/proteins. This problem can be addressed by refining the estimation using machine learning and artificial intelligence approaches, leveraging as training data publicly available experimental libraries, with gene expression profiling of cell lines stimulated with a variety of factors. This way, the compact matrix algebra of RFA would enable to train causal AI models, with direct molecular interpretation and well-defined empirical validation. Other possible directions of exploration are generalizations of the RFA model to take into account the distinction between inter- and intra-cellular regulation, as well as the differentiation between cell types.

Thus, we envision that RFA-based approaches will become prevalent and more sophisticated, in the pursuit of deeper understanding and more precise modeling of disease biology at molecular and cellular level. These efforts, hand in hand with extensive omics profiling, will facilitate rational drug design, with substantial improvements in the duration and success rates of drug projects.

6 Supplementary information

To compare prediction accuracy between the RFA and SFC algorithms, we used a dataset introduced in [50]. This dataset contains 66 different perturbations of a small network with 22 vertices, with each perturbation being performed under 200 secondary conditions, corresponding to different dose levels of EEGF and insulin, and different exposure times. Accordingly, the dataset contains measurements of $66 \times 200 = 13,200$ perturbation experiments. In each of those, we obtained the Regulation Flow with RFA and the Direction of Activity Change with SFC. The predicted values from both algorithms were used to calculate: a) mean accuracy, by comparing the sign of the prediction with the sign of the changes observed experimentally in genes/proteins (graph vectices); b) Spearman's correlation for the ordinal association between predicted and observed changes in genes/proteins. The results presented in Figure 5 illustrate that RFA and SFC have similar performance based on both metrics.

7 Code availability

An implementation of RFA is available at `https://github.com/AstraZeneca/regflow`.

References

[1] Research and development in the pharmaceutical industry. `https://www.cbo.gov/system/files/2021-04/57025-Rx-RnD.pdf`. Accessed: 2023-11-18.

[2] Katarzyna Smietana, Marcin Siatkowski, and Martin Møller. Trends in clinical success rates. *Nat Rev Drug Discov*, 15:379–380, 2016.

[3] Helen Dowden and Jamie Munro. Trends in clinical success rates and therapeutic focus. *Nat Rev Drug Discov*, 16:495–496, 2019.

[4] Paul Morgan, Dean G Brown, Simon Lennard, Mark J Anderton, J Carl Barrett, Ulf Eriksson, Mark Fidock, Bengt Hamren, Anthony Johnson, Ruth E March, et al. Impact of a five-dimensional framework on r&d productivity at astrazeneca. *Nature reviews Drug discovery*, 17(3):167–181, 2018.

[5] Denes Turei, Alberto Valdeolivas, Lejla Gul, Nicolas Palacio-Escat, Michal Klein, Olga Ivanova, Marton Olbei, Attila Gabor, Fabian Theis, Dezso Modos, Tamas Korcsmaros, and Julio Saez-Rodriguez. Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Molecular Systems Biology*, 2021.

[6] Guanming Wu and Robin Haw. Functional interaction network construction and analysis for disease discovery. *protein bioinformatics: from protein modifications and networks to proteomics*, pages 235–253, 2017.

[7] Clarivate metabase. `https://clarivate.com/`. Accessed: 2023-11-18.

[8] Qiagen knowledge base. `https://digitalinsights.qiagen.com/biomedical-knowledge-base/`. Accessed: 2023-11-18.

[9] Metaphacts. `https://metaphacts.com/`. Accessed: 2023-11-18.

[10] Evan O Paull, Daniel E Carlin, Mario Niepel, Peter K Sorger, David Haussler, and Joshua M Stuart. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 2013.

[11] Camille DA Terfve, Edmund H Wilkes, Pedro Casado, Pedro R Cutillas, and Julio Saez-Rodriguez. Large-scale models of signal propagation in human cells derived from discovery phosphoproteomic data. *Nature communications*, 6(1):8033, 2015.

[12] Glyn Bradley and Steven J Barrett. Causalr: extracting mechanistic sense from genome scale data. *Bioinformatics*, 33(22):3670–3672, 2017.

[13] Anika Liu, Panuwat Trairatphisan, Enio Gjerga, Athanasios Didangelos, Jonathan Barratt, and Julio Saez-Rodriguez. From expression footprints to causal pathways: contextualizing large signaling networks with carnival. *NPJ systems biology and applications*, 5(1):40, 2019.

[14] Robin Browaeys, Wouter Saelens, and Yvan Saeys. Nichenet: modeling intercellular communication by linking ligands to target genes. *Nature methods*, 17(2):159–162, 2020.

[15] Nikolaus Fortelny and Christoph Bock. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome biology*, 21:1–36, 2020.

[16] Ö Babur, A Luna, A Korkut, F Durupinar, MC Siper, U Dogrusoz, et al. Causal interactions from proteomic profiles: molecular data meets pathway knowledge. biorxiv. 2018, 2020.

[17] Andreas Krämer, Jeff Green, Jack Pollard Jr, and Stuart Tugendreich. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530, 2014.

[18] Daewon Lee and Kwang-Hyun Cho. Topological estimation of signal flow in complex signaling networks. *Scientific reports*, 8(1):5262, 2018.

[19] Daewon Lee and Kwang-Hyun Cho. Signal flow control of complex signaling networks. *Scientific Reports*, 9(1):14289, 2019.

[20] Carl D Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2023.

[21] Mustufa Jafry and Robert Sidbury. Rasopathies. *Clinics in dermatology*, 38(4):455–461, 2020.

[22] Marco Tartaglia, Ernest L Mehler, Rosalie Goldberg, Giuseppe Zampino, Han G Brunner, Hannie Kremer, Ineke van der Burgt, Andrew H Crosby, Andra Ion, Steve Jeffery, et al. Mutations in ptpn11, encoding the protein tyrosine phosphatase shp-2, cause noonan syndrome. *Nature genetics*, 29(4):465–468, 2001.

[23] Christian P Kratz, Giuseppe Zampino, Marjolein Kriek, Sarina G Kant, Chiara Leoni, Francesca Pantaleoni, Anne Marie Oudesluys-Murphy, Concezio Di Rocco, Stephan P Kloska, Marco Tartaglia, et al. Craniosynostosis in patients with noonan syndrome caused by germline kras mutations. *American Journal of Medical Genetics Part A*, 149(5):1036–1040, 2009.

[24] Martin Zenker, Denise Horn, Dagmar Wieczorek, Judith Allanson, Silke Pauli, Ineke Van Der Burgt, Helmuth-Guenther Doerr, Harald Gaspar, Michael Hofbeck, Gabriele Gillessen-Kaesbach, et al. Sos1 is the second most common noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *Journal of medical genetics*, 44(10):651–656, 2007.

[25] Bhaswati Pandit, Anna Sarkozy, Len A Pennacchio, Claudio Carta, Kimihiko Oishi, Simone Martinelli, Edgar A Pogna, Wendy Schackwitz, Anna Ustaszewska, Andrew Landstrom, et al. Gain-of-function raf1 mutations cause noonan and leopard syndromes with hypertrophic cardiomyopathy. *Nature genetics*, 39(8):1007–1012, 2007.

[26] Ion C Cirstea, Kerstin Kutsche, Radovan Dvorsky, Lothar Gremer, Claudio Carta, Denise Horn, Amy E Roberts, Francesca Lepri, Torsten Merbitz-Zahradnik, Rainer König, et al. A restricted spectrum of nras mutations causes noonan syndrome. *Nature genetics*, 42(1):27–29, 2010.

[27] Anna Sarkozy, Claudio Carta, Sonia Moretti, Giuseppe Zampino, Maria C Digilio, Francesca Pantaleoni, Anna Paola Scioletti, Giorgia Esposito, Viviana Cordeddu, Francesca Lepri, et al. Germline braf mutations in noonan, leopard, and cardiofaciocutaneous syndromes: molecular diversity and associated phenotypic spectrum. *Human mutation*, 30(4):695–702, 2009.

[28] Yoko Aoki, Tetsuya Niihori, Toshihiro Banjo, Nobuhiko Okamoto, Seiji Mizuno, Kenji Kurosawa, Tsutomu Ogata, Fumio Takada, Michihiro Yano, Toru Ando, et al. Gain-of-function mutations in rit1 cause noonan syndrome, a ras/mapk pathway syndrome. *The American Journal of Human Genetics*, 93(1):173–180, 2013.

[29] Guilherme Lopes Yamamoto, Meire Aguena, Monika Gos, Christina Hung, Jacek Pilch, Somayyeh Fahiminiya, Anna Abramowicz, Ingrid Cristian, Michelle Buscarilli, Michel Satya Naslavsky, et al. Rare variants in sos2 and lztr1 are associated with noonan syndrome. *Journal of medical genetics*, 52(6):413–421, 2015.

[30] Jennifer J Johnston, Jasper J van der Smagt, Jill A Rosenfeld, Alistair T Pagnamenta, Abdulrahman Alswaid, Eva H Baker, Edward Blair, Guntram Borck, Julia Brinkmann, William Craigen, et al. Autosomal recessive noonan syndrome associated with biallelic lztr1 variants. *Genetics in Medicine*, 20(10):1175–1185, 2018.

[31] Erin M Higgins, J Martijn Bos, Heather Mason-Suares, David J Tester, Jaeger P Ackerman, Calum A MacRae, Katia Sol-Church, Karen W Gripp, Raul Urrutia, and Michael J Ackerman. Elucidation of mras-mediated noonan syndrome with cardiac hypertrophy. *JCI insight*, 2(5), 2017.

[32] Yline Capri, Elisabetta Flex, Oliver HF Krumbach, Giovanna Carpentieri, Serena Cecchetti, Christina Lißewski, Soheila Rezaei Adariani, Denny Schanze, Julia Brinkmann, Juliette Piard, et al. Activating mutations of rras2 are a rare cause of noonan syndrome. *The American Journal of Human Genetics*, 104(6):1223–1232, 2019.

[33] Marialetizia Motta, Luca Pannone, Francesca Pantaleoni, Gianfranco Bocchinfuso, Francesca Clementina Radio, Serena Cecchetti, Andrea Ciolfi, Martina Di Rocco, Mariet W Elting, Eva H Brilstra, et al. Enhanced mapk1 function causes a neurodevelopmental disorder within the rasopathy clinical spectrum. *The American Journal of Human Genetics*, 107(3):499–513, 2020.

[34] Marialetizia Motta, Giulia Fasano, Sina Gredy, Julia Brinkmann, Adeline Alice Bonnard, Pelin Ozlem Simsek-Kiper, Elif Yilmaz Gulec, Leila Essaddam, Gulen Eda Utine, Ingrid Guarnetti Prandi, et al. Spred2 loss-of-function causes a recessive noonan syndrome-like phenotype. *The American Journal of Human Genetics*, 108(11):2112–2129, 2021.

[35] Diana Baralle, Chris Mattocks, Kamini Kalidas, Frances Elmslie, Joanne Whittaker, Melissa Lees, Nicola Ragge, Michael A Patton, Robin M Winter, and Charles ffrench Constant. Different mutations in the nf1 gene are associated with neurofibromatosis–noonan syndrome (nfns). *American journal of medical genetics Part A*, 119(1):1–8, 2003.

[36] Maria I Kontaridis, Kenneth D Swanson, Frank S David, David Barford, and Benjamin G Neel. Ptpn11 (shp2) mutations in leopard syndrome have dominant negative, not activating, effects. *Journal of Biological Chemistry*, 281(10):6785–6792, 2006.

[37] Tetsuya Niihori, Yoko Aoki, Yoko Narumi, Giovanni Neri, Hélène Cavé, Alain Verloes, Nobuhiko Okamoto, Raoul CM Hennekam, Gabriele Gillessen-Kaesbach, Dagmar Wieczorek, et al. Germline kras and braf mutations in cardio-facio-cutaneous syndrome. *Nature genetics*, 38(3):294–296, 2006.

[38] Pablo Rodriguez-Viciana, Osamu Tetsu, William E Tidyman, Anne L Estep, Brenda A Conger, Molly Santa Cruz, Frank McCormick, and Katherine A Rauen. Germline mutations in genes within the mapk pathway cause cardio-facio-cutaneous syndrome. *Science*, 311(5765):1287–1290, 2006.

[39] Hilde Brems, Magdalena Chmara, Mourad Sahbatou, Ellen Denayer, Koji Taniguchi, Reiko Kato, Riet Somers, Ludwine Messiaen, Sofie De Schepper, Jean-Pierre Fryns, et al. Germline loss-of-function mutations in spred1 cause a neurofibromatosis 1–like phenotype. *Nature genetics*, 39(9):1120–1126, 2007.

[40] Viviana Cordeddu, Elia Di Schiavi, Len A Pennacchio, Avi Ma'ayan, Anna Sarkozy, Valentina Fodale, Serena Cecchetti, Alessio Cardinale, Joel Martin, Wendy Schackwitz, et al. Mutation of shoc2 promotes aberrant protein n-myristoylation and causes noonan-like syndrome with loose anagen hair. *Nature genetics*, 41(9):1022–1026, 2009.

[41] Débora Bertola, Guilherme Yamamoto, Michelle Buscarilli, Alexander Jorge, Maria Rita Passos-Bueno, and Chong Kim. The recurrent ppp1cb mutation p. pro49arg in an additional noonan-like syndrome individual: Broadening the clinical phenotype. *American Journal of Medical Genetics Part A*, 173(3):824–828, 2017.

[42] Simone Martinelli, Alessandro De Luca, Emilia Stellacci, Cesare Rossi, Saula Checquolo, Francesca Lepri, Viviana Caputo, Marianna Silvano, Francesco Buscherini, Federica Consoli, et al. Heterozygous germline mutations in the cbl tumor-suppressor gene cause a noonan syndrome-like phenotype. *The American Journal of Human Genetics*, 87(2):250–257, 2010.

[43] Maria Lisa Dentici, Marcello Niceta, Francesca Romana Lepri, Cecilia Mancini, Manuela Priolo, Adeline Alice Bonnard, Camilla Cappelletti, Chiara Leoni, Andrea Ciolfi, Simone Pizzi, et al. Loss-of-function variants in erf are associated with a noonan syndrome-like phenotype with or without craniosynostosis. *European Journal of Human Genetics*, pages 1–10, 2024.

[44] Marwa Zafarullah and Flora Tassone. Molecular biomarkers in fragile x syndrome. *Brain Sciences*, 9(5):96, 2019.

[45] Verna Louhivuori, Annalisa Vicario, Marko Uutela, Tomi Rantamäki, Lauri M Louhivuori, Eero Castren, Enrico Tongiorgi, Karl E Åkerman, and Maija L Castrén. Bdnf and trkb in neuronal differentiation of fmr1-knockout mouse. *Neurobiology of disease*, 41(2):469–480, 2011.

[46] Nikolaj R Christensen, Jelena Čalyševa, Eduardo FA Fernandes, Susanne Lüchow, Louise S Clemmensen, Linda M Haugaard-Kedström, and Kristian Strømgaard. Pdz domains as drug targets. *Advanced therapeutics*, 2(7):1800143, 2019.

[47] Qiu-Lan Ma, Fusheng Yang, Sally A Frautschy, and Greg M Cole. Pak in alzheimer disease, huntington disease and x-linked mental retardation. *Cellular logistics*, 2(2):117–125, 2012.

[48] Galina Semenova and Jonathan Chernoff. Targeting pak1. *Biochemical Society Transactions*, 45(1):79–88, 2017.

[49] Mansuo L Hayashi, BS Shankaranarayana Rao, Jin-Soo Seo, Han-Saem Choi, Bridget M Dolan, Se-Young Choi, Sumantra Chattarji, and Susumu Tonegawa. Inhibition of p21-activated kinase rescues symptoms of fragile x syndrome in mice. *Proceedings of the national academy of sciences*, 104(27):11489–11494, 2007.

[50] Nikolay Borisov, Edita Aksamitiene, Anatoly Kiyatkin, Stefan Legewie, Jan Berkhout, Thomas Maiwald, Nikolai P Kaimachnikov, Jens Timmer, Jan B Hoek, and Boris N Kholodenko. Systems-level interactions between insulin–egf networks amplify mitogenic signaling. *Molecular systems biology*, 5(1):256, 2009.