A SIGNED ELEMENT-CONTRIBUTION FRAMEWORK FOR DESCRIBING THE FUNCTION OF FUSION PROTEINS

Katsuhiko Murakami
Fujitsu Research, Fujitsu Ltd.
Kawasaki, 211-8588, Japan
E-mail: murakami.kt@fujitsu.com

A substantial fraction of cancers are driven by fusion genes, and interpreting a newly observed fusion candidate remains a practical problem. What a fusion protein does is decided by which functional elements survive the breakpoint, so its function is not determined by the presence or absence of domains, and existing annotation resources are insufficient for inference. What they omit is direction: whether keeping a given element moves the phenotype toward disease or away from it. We formalize the classes of oncogenic mechanism as a sum of context-dependent signed element responses, extend the annotation of fusion protein elements with that direction, and catalogue the conditions under which it reverses. Applying the framework to solid-tumour fusions from COSMIC selected for breakpoint diversity, we build a tool that matches an unknown fusion against known ones on a role composition expressing the role of each element in a pathogenicity context, and that declines to answer, with a coverage guarantee, when no known mechanism is close enough. Identification of the mechanism class was not achieved, and we locate the reason in the representation itself: between breakpoints of one gene pair what varies is how much protein remains, not which functional roles remain. We report term by term which parts of the framework current knowledge and current annotation can support.

Keywords: Gene fusion; functional annotation; conformal prediction; breakpoint; oncogenic mechanism.

1. Introduction

Computational prediction of protein function has long relied on homology transfer. Recent AI methods have greatly improved the detection of distant homologues, yet determining protein function remains difficult when the functional elements are known but their organization is altered.

Fusion proteins illustrate this limitation, and the problem is a practical one: sequencing now reports new fusion candidates whose consequence must be judged. Similarity search usually identifies their retained domains correctly, but function depends on which elements survive the breakpoint and how they act after rearrangement. Breakpoint variants of the same gene pair can differ in disease and in drug response, as BCR-ABL p190 and p210 do, and as EML4-ALK variants do under tyrosine kinase inhibitors. Homology provides a parts list but not the consequence of the rearrangement.

Existing resources describe the parts but not their direction. Pfam, InterPro and Gene Ontology annotate domains and molecular functions; fusion-specific resources record domain retention [1-3] or predict pathogenicity for a gene pair or a transcript [4-10]; none states whether keeping a given element pushes the phenotype toward disease or away from it, which we call the *sign* of the element. Gene Ontology cannot supply it, because it annotates gene products rather than elements [11] and its negative-regulation terms relate processes rather than the consequence of losing an element.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

Reviews of fusion activation describe the mechanisms [12], and one already links the intron in which the break falls to the retained domains and to disease in ABL1 fusions [13], but neither assigns a direction to each element.

Our own earlier work [10] predicted the pathogenicity of a fusion as a whole and could not say which retained element carries the effect; the present framework is a response to that limitation, on an analysis set rebuilt independently from COSMIC that does not overlap the earlier data. Since no published method returns a signed description at the level of the element, we state these differences rather than report a head-to-head comparison.

A description is worth having if it separates what differs, so writing down the sign raises two questions. Does the set of signed elements that survive a breakpoint distinguish the mechanism by which a fusion is oncogenic? And when the breakpoint moves within one gene pair, as between the variants above, what changes: which roles are kept, or only how much of the protein? The answer to the second question turns out to explain the answer to the first.

We formulate the contribution of each element as a context-dependent signed response, catalogue the conditions under which the sign reverses, and build a precedent search on the resulting role composition, calibrated by conformal prediction so that it declines to answer when nothing known is close enough. We then report, term by term, which parts of the framework current evidence supports, which it rejects, which cannot be tested for want of annotation, and which are defined but not yet built.

2. Theory: signed description under a structural event

The framework below does not aim to exceed existing methods in predictive accuracy but to bring observations reported separately into a common descriptive form. We therefore keep the components deliberately simple: role assignment is rule-based without learning, the precedent search uses one-nearest-neighbour with Euclidean distance, and its reliability rests on the distribution-free guarantee of conformal prediction (Section 3.6). Keeping the prediction simple is what allows the location of the discriminating information to be traced by hand.

2.1. *Objects and operations*

We use *functional elements* to mean annotation units that contribute to protein function, including domains, regions and motifs. A fusion joins two genes, which we call the *gene pair*; the position at which the two proteins are joined is the *junction*, and two junctions of the same gene pair are *breakpoint variants*. The *mechanism class* is the type of oncogenic mechanism a fusion belongs to, and we use three: kinase activation, transcription-factor chimera, and other. Let a molecule M consist of a set of annotated functional elements $E = \{e_1, \ldots, e_n\}$. A structural event τ transforms M into M'. For each element we define a *functional integrity* $x_i \geq 0$, where $x = 1$ means intact, $x = 0$ means lost, $0 < x < 1$ means partial retention, and $x > 1$ means duplication. The retained fraction in a fusion is a special case of x. We do not cap x at 1 because events such as internal tandem duplication increase the number of copies of an element.

2.2. *Signed response model*

We define the annotation as three maps. A *role map* ρ assigns each element a role, written r_i for element i, among effector, autoinhibitory, oligomerization and neutral. A *sign map* s returns -1, 0 or +1 given a role and a context c. A *response function* g converts functional integrity into the magnitude of the contribution. The estimated contribution of role r to the phenotype of M', written $\hat{y}_r(M')$, is

$$\hat{y}_r(M') = \frac{1}{|E_r|} \sum_{i:r_i=r} s(r,c)\, w_i\, g_r(x_i)\,, \qquad E_r = \{i : r_i = r\} \tag{1}$$

where E_r is the set of elements with role r and $w_i \geq 0$ is the weight of the element. In this paper all weights are equal ($w_i = 1$) and g is the identity, so the estimator reduces to the mean retained fraction within each role; Eq. (1) is a definition, and no parameter is fitted to data. An effector acts toward pathogenicity when retained and takes sign +; an autoinhibitory element acts toward pathogenicity when lost and takes sign -; a neutral element takes 0. What existing fusion annotation resources do is the case in which s is always +1 and g is one if any part of the element remains and zero otherwise: an unsigned, all-or-nothing judgment, a degenerate case of this framework. Our generalization has three parts: introducing the sign s, making the response g continuous, and letting the sign depend on context.

The advantage of this framework is that each term can be tested separately as a comparison of nested models: whether the first-order term is supported, whether a non-monotone term adds explanatory power, and whether separating the sign by role is worthwhile. The same framework applies to other events that change the composition of elements, such as exon skipping (MET exon 14), in-frame deletion (EGFRvIII) and internal tandem duplication (FLT3-ITD).

2.3. *Two-stage resolution of the sign and reversal conditions*

The sign is resolved in two stages. The default is a map at the level of the domain family; if an entry exists in the catalogue of reversal conditions at the level of the gene or of the gene pair, that entry takes precedence. The structure is that of a position-specific scoring matrix overriding a general substitution matrix at particular positions.

Besides the family default, reversal conditions arise at the level of the gene, such as the PKC catalytic domain or the kinase domains of tumour-suppressive kinases, and of the gene pair, such as CLTC; Section 4.3 lists them.

If no entry applies and there is no assurance that the family default holds for the gene in question, no sign is assigned and the element is left unresolved, as the precedent search returns the empty set when nothing known is close enough.

The catalogue of reversal conditions is keyed on genes and gene pairs, not on fusions. A new combination of known genes can therefore receive the sign override even though the fusion itself is unknown. In the current implementation, however, gene-level resolution is not connected, so this route is defined in the framework but not reached (Section 4.10).

3. Materials and methods

3.1. *Data and analysis set*

We used COSMIC Fusion v100 (GRCh37) [14] with gene and transcript models from Ensembl GRCh37 release 73 [15] accessed through pyensembl. We restricted the data to solid tumours and, where both orientations of a fusion were reported in the same specimen, normalized to the driver orientation, obtaining 224 pairs for which the reading frame of the fusion protein could be reconstructed. We then restricted these to the 82 pairs in which two or more distinct predicted protein sequences were observed. Fifteen of those pairs are promoter-hijacking pairs, in which the 5' gene supplies only regulatory sequence and no protein; two of them (CDH11-USP6 and COL1A1-USP6) were removed before annotation and the remaining thirteen at the mechanism-class step, so the counts pass through 80 annotated pairs with 457 junctions and end at 67 pairs and 373 junctions, the main analysis set. This selection is deliberate: a change between breakpoints of one pair is defined only when the pair has more than one junction, so the 142 of 224 pairs (63%) seen with a single junction are excluded and the results do not generalize to them.

3.2. *Definition of the junction*

The proteins encoded by the two fusion partner genes are called the 5' and 3' partner proteins, and a junction is the position at which they are joined. Because the count depends on what is treated as the same junction, we compared three definitions (experiment E-1): agreement of the text representation, agreement of the pair of genomic coordinates, and agreement of a hash of the predicted protein sequence. We adopt the third. Since the question is the function of the fusion protein, coordinate differences that give the same protein are functionally equivalent. For the same reason, fusions with different 5' partners that yield the same protein are merged; among the 80 annotated pairs this happens for three junctions, all of the promoter-hijacking type, and it does not occur in the 67-pair analysis set.

3.3. *Role annotation, retained fraction, and the catalogue of reversal conditions*

For the 5' and 3' partner proteins of the 80 annotated pairs, the feature table of UniProt [16] lists 9,081 functional elements over 108 proteins. To assign a role and an expected sign to these 9,081 elements we constructed rules over element names (experiment E-2). Two element counts therefore appear in this paper: the 9,081 elements of the 80 annotated pairs, used for the rule table and for the null comparison of Section 4.4, and the 8,131 of them that belong to the 90 proteins of the 67-pair analysis set, used for the GO survey, the precedent search, the within-pair tests and the tables of Sections 4.8 and 4.9. The complete set of rules is given in Table 1.

We mapped the genomic coordinate of the breakpoint onto the protein coordinate of the representative transcript and computed, for each functional element (domain, region, motif and the like), the fraction retained in the fusion protein, which we call the *retained fraction*. This computation depends only on coordinates and not on the reading frame or the position of the stop codon. The *role composition* of a junction is a four-dimensional vector of the mean retained fraction for each of the four roles.

The catalogue of reversal conditions was assembled from primary literature describing mechanisms in which the expected sign does not hold. Each row records the target element, the default interpretation, the condition under which reversal occurs, and the mechanism.

3.4. *Survey of existing resources and of GO*

To ask whether existing resources can supply a sign at junction resolution, we tabulated the main fusion-related resources by whether they reach breakpoint resolution, whether they carry a pathogenicity label, and how they treat fusions (experiment E-3).

To test whether GO can supply the sign, we retrieved GO annotations for the 90 partner proteins from UniProt. GO annotations were retrieved with and without electronically inferred (IEA) annotations, to test whether the conclusions depended on automatically assigned annotations. We extracted every term whose name contains "negative regulation" and classified whether the target is the activity of the annotated protein itself or another pathway. Classification was not automated: the rule was fixed before the results were seen and every term was read.

3.5. *Null comparison for the distribution of breakpoints*

To ask whether breakpoint positions are random, we constructed a null model in which breakpoints fall uniformly on the protein coordinate (experiment E-4) and compared, against the observed data, the probability that an element is completely retained and the probability that it is completely lost. Confidence intervals were obtained by bootstrap resampling at the level of the gene pair and by the Clopper-Pearson method [17]. Because real breakpoints fall in introns rather than at arbitrary amino-acid positions, we repeated the comparison with two further null models in which the cut is placed at an intron of the representative transcript, either uniformly over the introns or with probability proportional to intron length, the latter being the same structural assumption used for the selection strength. Point estimates are means over elements, and it is the bootstrap that resamples gene pairs.

3.6. *Calibrated precedent search*

The purpose of experiment E-5 is to return known fusions whose role composition is close to a new junction, and to state in a calibrated form how close is close enough to be trusted. The E-value of BLAST gives the probability of such similarity arising by chance; it does not give the probability that a hit carries the correct function. What is needed is a guarantee of the latter kind.

We therefore used conformal prediction [18,19]. It takes any score as a *nonconformity score*, a measure of poorness of fit, and returns a candidate set of mechanism classes. For a query x and a candidate class y, a conformal p-value is defined against the nonconformity scores $A_1, \ldots, A_n$ of a calibration set as

$$p_y(x) = \frac{|\{i: A_i \geq A(x,y)\}| + 1}{n+1}, \qquad \Gamma_\alpha(x) = \{y \in Y: p_y(x) > \alpha\} \tag{2}$$

where Y is the set of mechanism classes, n is the size of the calibration set and α the target error rate, so that, as far as calibration data and query data are exchangeable, $P(y_{\text{true}} \in \Gamma_\alpha(x)) \geq 1 - \alpha$ holds. The guarantee depends neither on the quality of the score nor on the learning model. Unlike models that produce uncalibrated point predictions and offer no guarantee on future data, conformal

prediction returns statistically backed sets for a target error rate without retraining, which matches our motivation: evidence with a mathematical basis rather than fluctuating uncertainty. A larger set means greater uncertainty, and an empty set states that nothing known is close enough.

We used, as the nonconformity score, the minimum Euclidean distance from the role composition of the query junction to known junctions belonging to each mechanism class, and returned a candidate set of mechanism classes ($\alpha = 0.2$, nominal coverage 80%). If the query retrieves another junction of the same gene pair, the match is made by the shortcut of the gene name rather than by role composition; to avoid this we removed the entire pair to which the query belongs from the search set (leave-one-pair-out, 20 random seeds). Coverage is the fraction of queries for which the returned set contained the true mechanism class.

3.7. *Test of the within-pair effect and variance partition of the role composition*

To test whether the strength of selection changes when the breakpoint moves within a pair, we used a mixed-effects model absorbing the gene pair as a random effect (experiment E-6), comparing by likelihood-ratio test a model with only first-order monotone terms against one adding non-monotone terms. The response variable is the selection strength of a junction on the log scale, log(O/E), where O is the number of specimens in which the junction was observed and E is the number expected under a structural null in which a breakpoint falls in an intron with probability proportional to intron length. The monotone model has four predictors, the mean retained fraction of the domain and region layers on each side; the role-split model also has four, the retained fraction of the four roles. The two are not nested in one another and each is compared against the same null model.

If nothing is detected in that test, the question remains whether there was any variation within pairs to predict. To measure this without using the response variable, we partitioned the variance using the predictors alone. For each role we computed the fraction of pairs in which the retained fraction crosses the functional boundary of 0.5 between junctions of the same pair, and call this the *crossing rate*; 0.5 marks whether an element is functionally preserved or lost. The crossing rate of the neutral component was taken as the baseline and the decision rule was fixed before the analysis was run. A high crossing rate means that the role varies within a pair, so the annotation resolves differences between breakpoints of the same gene pair; a low crossing rate means the role is effectively a property of the pair and the annotation cannot express such differences. To check whether the crossing rate depends on the number of junctions, we computed crossing rates stratified by junction count, applied direct standardization to the junction-count distribution of the neutral component, and used promoter-hijacking pairs as a negative control.

3.8. *Ceiling of annotation and matching of clinically distinguished variant pairs*

To measure how far annotation could reach if it were extended, we scanned the UniProt feature table of the 90 partner proteins and counted whether a description corresponding to each role exists (experiment E-7), comparing the current number of annotations with the number of proteins for which such a description exists, which we take as the ceiling. For three variant pairs reported to show different clinical outcomes, we computed the mean retained fraction by role and asked whether the two members can be distinguished at the level of roles.

Data and code availability. The analysis set is derived from COSMIC, whose licence does not permit redistribution, so the junction-level data cannot be released; the rules that assign roles are listed in full in Table 1 and the catalogue of reversal conditions in Section 4.3, and the steps above are described in enough detail to be reimplemented from COSMIC and UniProt obtained from their providers. Analyses use the releases given in Section 3.1 and UniProtKB retrieved in July 2026. Analyses were performed with Python 3.14, SciPy 1.18 and statsmodels 0.14.

4. Results

4.1. *Neither existing resources nor GO supplies the sign*

The survey showed that although resources reach breakpoint resolution, none annotates pathogenicity experimentally at that resolution. COSMIC reaches the junction but its label is only the observed case count; FusionGDB2 annotates breakpoints and domain retention but only computationally. OncoKB [20], CIViC [21] and ChimerDB carry pathogenicity labels but at the level of the gene pair, without distinguishing junctions. The ClinGen/CGC/VICC standard operating procedure, the community rule set for calling a somatic variant oncogenic [22], places fusions out of scope. Resources that reach the breakpoint carry no label, and resources that carry a label do not reach the breakpoint.

In the GO survey, of the 118 "negative regulation" terms carried by the 90 proteins, one (0.8%) had the activity of the annotated protein itself as its target and 115 (97.5%) targeted other pathways or processes; the remaining two could not be decided from the term name alone (negative regulation of protein-containing complex assembly on RAF1, and negative regulation of catalytic activity on NOTCH1), and rather than force the binary rule we left them unassigned. The term on BRAF, "negative regulation of apoptotic process", is of the latter kind; suppression of apoptosis is oncogenic, so the direction of the string and the direction of the contribution to the phenotype are opposite. GO's negative regulation describes a relation between molecules, a quantity distinct from the direction an element carries under a structural event.

4.2. *Content and coverage of the role annotation*

Table 1 shows the result of applying the rules to the 9,081 elements (experiment E-2). A clearly non-neutral role could be assigned with confidence to about 9% of the total; the remainder fell to Neutral (default). The mean retained fraction of 0.000 for autoinhibitory elements means that every observed fusion has lost autoinhibition. The neutral component is in effect a proxy for the fraction of sequence retained in the fusion protein.

4.3. *Catalogue of conditions under which the sign reverses*

The catalogue starts from the two processes that reviews of fusion activation agree on, enforced oligomerization and inactivation of inhibitory domains [12], asks of each where the expected sign can fail, and anchors every type to a primary report; it is a curated list of counterexamples rather than a systematic review, and claims neither completeness nor frequency. We collected seven types of mechanism in which the expected sign does not hold; expanded to concrete fusion instances these

become eleven, three of which can be confirmed in our analysis set and the others rest on the literature. Against a default of "should be lost", the N-terminal autoinhibition of BRAF can be retained and the protein still be oncogenic, because the 5' partner tethers the near-full-length kinase to the membrane system; an intact dimer interface is required, but it is not what the partner supplies [23]. The PKC regulatory domain gives dominant-negative suppression when the kinase is lost [24], and the ABL cap is layered, so the myristoyl cap is lost while SH3 and SH2 survive and impose a second level of autoinhibition [25,26]. Against a default of "should be retained", the PKC catalytic domain does not confer stable activity [24], the beta-propeller of EML4 breaks folding when partially retained but is stable when completely lost [27], and in CLTC the region annotated as the trimerization domain is retained in CLTC-ALK but lost in CLTC-TFE3, whose partner carries a dimerization motif of its own [28,29]; we read the shift of oligomerization from the 5' to the 3' side off the retained regions, as neither fusion has been assayed for oligomerization, and it is a pair-level condition. Two entries fall outside these defaults: the HMGA2 3'UTR, whose loss escapes let-7 repression [30] and which lies outside the protein role vocabulary, and kinase domains in general, since in tumour-suppressive kinases such as STK11, CSK, MATK and DAPK1 the family does not determine the sign. All four BRAF fusions in our set had lost the autoinhibitory region, consistent with the known mechanism; the reversal cases reported in the literature are not present in our set.

Table 1. Rules for role assignment and their distribution in the data.

Role (expected sign)	Elements	Junctions	Mean retained fraction	Matching rule (applied in this order)
Effector (+)	418	319	0.636	name contains kinase; or ETS, basic-leucine, bHLH, homeobox, zinc finger, DNA, RanBP2, bromo, SET, PHD, chromo, Myb
Autoinhibitory (-)	44	22	0.000	name contains RBD, Ras-binding, phorbol or autoinhibitory, or the word C1
Oligomerization (+)	389	149	0.629	coiled-coil layer; or name contains SAM, Pointed, PB1, SCAN, trimerization, oligomerization, zipper, or begins with WD
Neutral (explicit)	2,481	451	0.547	disordered or motif layer; region not matched above
Neutral (default)	5,749	454	0.702	domain not matched by any rule above (default)

4.4. *Breakpoints converge on effector retention and autoinhibition loss*

To ask whether breakpoint positions are non-random, experiment E-4 compared a uniform-cut null model against the observations. The retained fraction by role is bimodal (Fig. 1A): 61% of effector elements are completely retained and 36% fall below 0.5, with almost nothing in between. Against the null, complete retention of the effector is 1.35-fold (95% CI 1.21-1.51) and complete loss of autoinhibition is 1.47-fold (1.41-1.49), both with lower bounds above one; autoinhibition is lost completely in all five pairs where it is defined, so that interval reflects only the uncertainty of the null. The probability that both hold simultaneously is 4.29-fold that of the null. The median case count of the group not retaining the effector is 2, outside the typical composition of recurrent fusions. Placing the cut at an intron instead makes the departure smaller but does not remove it: complete retention of the effector is 1.27-fold (1.11-1.42) when cuts fall uniformly over introns and 1.20-fold (1.10-1.28) when they are weighted by intron length, and complete loss of autoinhibition is 1.35-

fold (1.32-1.36) and 2.79-fold (2.74-2.97). Exon structure thus accounts for part of the departure, and the lower bound stays above one under every null.

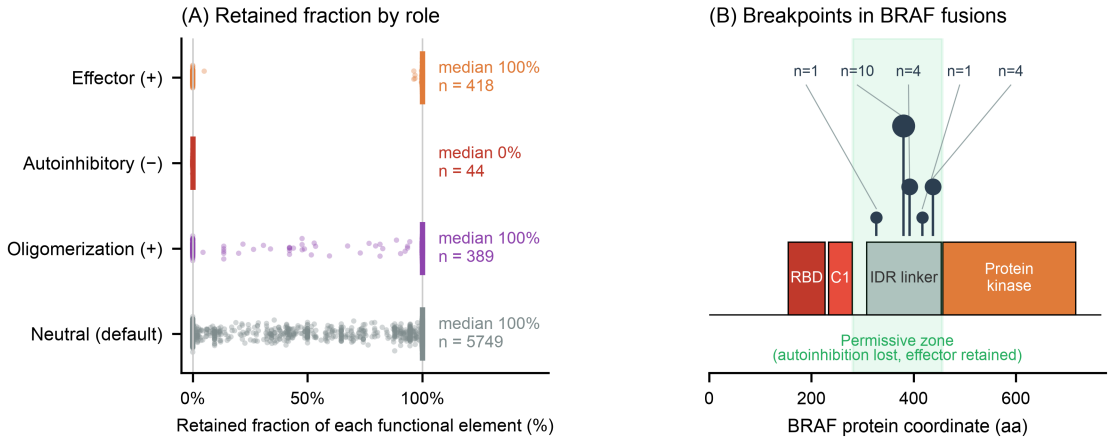


Fig. 1. (A) Retained fraction of each functional element by role, from 0% (lost entirely) to 100% (kept entirely). Each point is one element and the thick bar marks the median. Effector and autoinhibitory elements are almost always kept or lost whole, oligomerization elements mostly so, and partial cuts fall mainly on neutral elements. (B) Breakpoints in BRAF fusions on the amino-acid coordinate of BRAF. Boxes are domains (red, autoinhibitory; orange, effector kinase; grey, intrinsically disordered linker); the green band is the permissive zone, where autoinhibition is lost and the effector retained. Each stem marks an observed breakpoint, with circle size and the number along the top giving its junction count. In every fusion observed here BRAF is the 3' partner (20 junctions in four pairs), so the N-terminal autoinhibitory region is the part that can be lost; fusions with BRAF as the 5' partner are known but do not pass the filters used here.

All 20 junctions of BRAF fusions fall inside the region that loses autoinhibition and retains the effector kinase (Fig. 1B). This is a case study confined to BRAF rather than a general validation, since every pair for which autoinhibition is defined belongs to the BRAF/RAF1 family.

4.5. *Precedent search returns a guaranteed set*

In experiment E-5 (Section 3.6), observed coverage was 0.778, close to the nominal 0.80; at coverage matched to 80%, the mean size of the non-empty sets was 1.57, below the frequency-ordered baseline of 2.00, with an information gain of 0.80 bits. Removing the query's own gene pair is necessary: without it, the nearest neighbour would come from the same pair for 202 of 373 junctions (54%). Other pairs that share one gene with the query do remain in the search set, however, and coverage is higher for queries with such a pair (0.812) than for those without (0.566), so part of the search's performance may rest on a shared partner gene rather than on the role composition alone. For 15.9% of queries the set was empty: nothing known was close enough. To ask whether the roles carry information beyond the retained fraction, we repeated the search with the role composition replaced by the total retained fraction alone. Coverage is unchanged, since conformal prediction calibrates it by construction, but the returned sets are larger: 2.04 classes at matched coverage against 1.57 for the role composition, which is the size of the frequency-ordered baseline of 2.00, and the information gain falls from 0.80 to 0.50 bits. Permuting the role labels among elements gives 2.17 classes and 0.39 bits, so the gain is not an effect of the number of dimensions. Replacing each junction by the mean role composition of its gene pair left the search essentially unchanged (mean

set size 1.53 versus 1.57), consistent with mechanism class being a pair-level label in the present evaluation. The discriminating quantity is an amount, but averaging that amount over all elements destroys it; the roles are what make it informative.

4.6. *Mechanism-class identification does not succeed, but the cause is located*

The nearest neighbour had the same mechanism class as the query more often than always answering the most frequent class would (0.68 against 0.59), but not consistently: it did better in 15 of 20 splits and worse in 5, and the difference is not distinguishable from zero. Coverage was uneven across classes, at 0.593 for kinase activation.

The cause can be identified. Discrimination by the role composition is carried in effect by the neutral component, the total amount of sequence retained in the fusion protein. In oncogenic fusions the three components effector, autoinhibitory and oligomerization converge on approximately (1, 0, 0) and produce no difference between junctions. In the three cases KIAA1549-BRAF, SND1-BRAF and MYB-NFIB these three components are identical, and the ranking is determined only by small differences in the neutral component (distances 0.0006 and 0.0045, Fig. 2). In Fig. 2 these distances appear as 0.00: the compositions differ only beyond the second decimal, in the neutral component. The role space is not completely collapsed only because the neutral component is continuous; on the three functional components alone, many junctions coincide at the same point. The information the current role vocabulary retains is therefore how much was retained, and not which effector, which is what distinguishing mechanism classes would require.

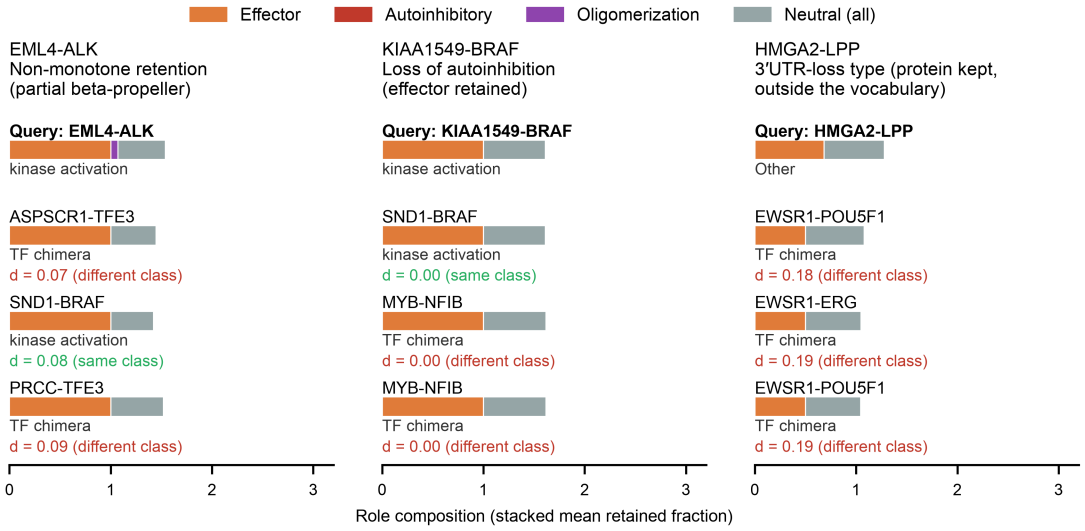


Fig. 2. Precedent search for three query junctions. The heading of each panel names the query and the type of case it illustrates. Each row is one junction, and the bar shows its role composition: the four segments are the mean retained fraction of the effector (orange), autoinhibitory (red), oligomerization (purple) and neutral (grey) elements, drawn end to end. The first row of each panel, in bold, is the query; the three rows below it are the nearest known junctions returned by the search, ordered by Euclidean distance in that four-dimensional space. Under each row are its mechanism class and its distance d from the query, printed in green when the class matches the query's and in red when it does not; d is rounded to two decimals, so 0.00 means a distance below 0.005. Junctions of the query's own gene pair are excluded from the search, so the neighbours always come from other pairs.

4.7. *A monotone trend exists within pairs, but no information beyond it*

In experiment E-6, adding the retained fraction as an explanatory variable improved the fit significantly ($\chi^2 = 46.8$, $p = 5.0 \times 10^{-6}$), and the first-order monotone term alone was significant ($\chi^2 = 35.3$, 4 df, $p = 4.1 \times 10^{-7}$). Within a pair, the more functional elements are retained, the more likely the junction is to be selected. Splitting the terms by role reaches a smaller statistic than the averaged monotone term at the same degrees of freedom ($\chi^2 = 33.7$ against 35.3); the two models are not nested in one another and each is compared against the same null, so the role split adds no explanatory power over the average.

Information beyond the monotone trend, that is, the non-monotone information carried by the breakpoint being at an intermediate position, was not detected ($\chi^2 = 11.5$, 8 df, $p = 0.17$). This is not for want of instances: 44 junctions across 14 pairs carry partially retained oligomerization elements.

4.8. *The role composition is a pair-level property on the decisive axes*

The variance partition of experiment E-6 (Table 2) asks whether there was within-pair variation to predict at all. The two mechanistically decisive axes, effector retention and autoinhibition loss, are almost constant within a pair. In 51 of 55 pairs (92.7%) the effector shows no difference crossing the functional boundary between junctions of the same pair, and the median within-pair standard deviation is 0.000. Autoinhibition shows no crossing in any of the five testable pairs; with 0 of 5 the exact binomial (Clopper-Pearson) two-sided 95% interval is [0, 0.522], so a true crossing rate as high as 0.522 cannot be rejected and this axis has no power at the present sample size.

The only functional axis that varies within a pair is oligomerization, whose crossing rate of 0.385 exceeds the neutral baseline of 0.149. We confirmed that this comparison is not distorted by sensitivity to junction count. The baseline increases monotonically with junction count, but the junction-count distributions of the three roles almost coincide and direct standardization leaves the crossing rates essentially unchanged (effector 0.073 to 0.074; oligomerization 0.385 to 0.390). The ordering of effector below neutral holds in all four strata, and the negative control gave 0.000 for both effector and autoinhibition.

Table 2. Variance partition of the role composition (67 pairs, 373 junctions).

Role	Pairs testable	Between-pair SD	Median within-pair SD	Crossing rate	Standardized
Effector	55	0.388	0.000	0.073 (4/55)	0.074
Autoinhibitory	5 (no power)	0.000	0.000	0.000 (0/5)	-
Oligomerization	26	0.304	0.011	0.385 (10/26)	0.390
Neutral (baseline)	67	0.245	0.040	0.149 (10/67)	0.149

4.9. *The ceiling of annotation lies on the resource side*

Against the UniProt ceiling (experiment E-7), the effector is at 55 pairs against a ceiling of 55, or 100%, and oligomerization at 26 against 28, or 93% (Table 3). Only autoinhibition is below, at 5 against 8, or 63%, and for the remaining 59 pairs there is no description in UniProt at all. None of ALK, RET, ROS1 or NTRK1 carries a description corresponding to autoinhibition. Juxtamembrane

autoinhibition in receptor tyrosine kinases is established biology, but it does not exist as a machine-readable annotation.

Table 3. Current role annotation against the UniProt ceiling.

Role	Current annotation	UniProt ceiling	Current/ceiling	No description in UniProt
Effector	55 pairs	55 pairs	100%	12 pairs
Oligomerization	26 pairs	28 pairs	93%	39 pairs
Autoinhibitory	5 pairs	8 pairs	63%	59 pairs

Of the three variant pairs reported to differ clinicopathologically, only EML4-ALK differed at the level of roles, and the difference appears on the oligomerization axis, the only axis that varies within pairs (Section 4.8). Variant 1 (E13;A20) keeps part of the EML4 beta-propeller, which leaves the fold incomplete and the protein unstable, whereas variant 3a/b (E6;A20) loses that region entirely and is more stable; the stability difference is reported to track the response to ALK inhibitors [31], and our retained fractions on this axis, 0.429 against 0.071, follow it. The region of EML4 that drives oligomerization is retained in every variant [32], so the difference along this axis reflects how much of the rest of EML4 survives, which bears on the stability of the protein rather than on oligomerization itself. The role-level difference appears in this pair because an element with a non-monotone response is partially retained. In KIAA1549-BRAF the roles on the BRAF side are identical between variants (effector 1.0, autoinhibitory 0.0) and the difference appears only in the retained length on the 5' side (neutral 0.75 against 0.65); in NAB2-STAT6 no functional role is assigned at all (neutral 0.724 against 0.835), because the DNA binding and transcriptional activation of STAT6 have no corresponding concept in the current vocabulary. Both of these pairs are reported to differ in tumour location or clinicopathological phenotype, but neither difference is an established independent predictor of survival or recurrence [33-36] and the vocabulary cannot express either of them.

4.10. *How far each term of the framework is supported*

Table 4 organizes how far each term of the framework is supported by current knowledge and resources. The reasons for lack of support fall into three kinds. Two further terms are stated here rather than in the table: gene-pair-level context is partly implemented (Section 4.3), and the neutral term is a diagnosis rather than a test (Section 4.6).

Table 4. Audit of the model terms.

Term	Verdict	Evidence and caveats	What is needed
First-order term (effector)	Supported	1.35-fold departure from the uniform null (1.21-1.51) (Section 4.4); likelihood-ratio test chi-square 46.8 (Section 4.7)	-
First-order term (autoinhibitory)	Partly supported	Complete loss 1.47-fold (1.41-1.49) (Section 4.4); no power within pairs at five pairs (Section 4.8)	Annotation of autoinhibition; absent from UniProt for 59 pairs (Section 4.9)
Non-monotone term	Tested, not supported	chi-square 11.5, 8 df, p = 0.17; 44 junctions across 14 pairs available (Section 4.7)	A fifth role (structural integrity), or subdivision of the effector
Sign split by role	Tested, not supported	chi-square 33.7 against 35.3, no gain at equal df (Section 4.7)	As above
Context at gene level	Not implemented	0 of 8,131 elements resolved (0.0%); a prototype exists but is not connected	Implementation only; no additional knowledge required

5. Discussion

Three independent analyses (Sections 4.6-4.8) agree: breakpoint variants of the same gene pair differ mainly in the amount of protein retained, not in the role composition. Effector retention and autoinhibition loss are largely fixed within a pair, and oligomerization is the only functional axis that often varies, which is why variants of EML4-ALK can be told apart [31].

The negative result is informative. Mechanism-class identification did not exceed the baseline because the present vocabulary records whether functional roles are retained, but not which specific effector performs the function. The representation therefore preserves quantitative variation in retained protein length more effectively than qualitative differences among effectors. This identifies the principal direction for extending the framework: subdividing broad functional roles, particularly the effector category, into biologically distinct subclasses.

The framework also records where current knowledge stops: autoinhibitory regions in particular remain under-annotated in UniProt, which limits the coverage of role annotation (Section 4.9).

Clinical variant interpretation separates evidence in a similar way: the ACMG/AMP framework assigns direction and strength to individual variants [37], and the present framework applies the same principle to functional elements under structural alteration, including alterations other than fusions, such as protein-truncating variants.

The present study has several limitations. The autoinhibitory axis is supported by only five informative gene pairs, convergence on effector retention may partly reflect ascertainment bias in reported oncogenic fusions, protein coordinates differ between UniProt and Ensembl for a subset of genes, and the evaluation is restricted to recurrent gene pairs. A benchmark against experimentally characterized fusion proteins would need pathogenicity labels at breakpoint resolution, which no current resource carries (Section 4.1); assembling them junction by junction is the work this framework points to.

6. Conclusion

Whether a fusion protein drives cancer depends on which of its parts survive the breakpoint and on the direction in which each surviving part pushes the protein. We wrote that direction down as a context-dependent sign on each element; annotation that records only whether a domain is present is the special case in which every sign is positive. Applied to 67 gene pairs and 373 junctions, the framework returns known fusions with a stated error rate and declines to answer when nothing known is close enough.

It does not identify the mechanism class, and it shows why: variants of one gene pair keep the same effector and the same loss of autoinhibition and differ mainly in how much protein remains, so the class is a property of the gene pair and the breakpoint acts below it. Each term of the framework is marked as supported, rejected, untestable for want of annotation, or not yet built, and the next step this points to is functional labels at junction resolution rather than another predictor.

References

1. C. Murphy and O. Elemento, AGFusion: annotate and visualize gene fusions, https://github.com/murphycj/AGFusion (accessed 2026).

2. P. Kim, H. Tan, J. Liu, H. Lee et al., FusionGDB 2.0: fusion gene annotation updates aided by deep learning, Nucleic Acids Res. 50, D1221 (2022).
3. H. Kumar, L.-Y. Tang, C.-C. Yang and P. Kim, FusionPDB: a knowledgebase of human fusion proteins, Nucleic Acids Res. 52, D1289 (2024).
4. M. Shugay, I. Ortiz de Mendíbil, J. L. Vizmanos et al., Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions, Bioinformatics 29, 2539 (2013).
5. F. Abate, S. Zairis, E. Ficarra et al., Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer, BMC Syst. Biol. 8, 97 (2014).
6. M. Lovino, M. Montemurro, V. S. Barrese and E. Ficarra, Identifying the oncogenic potential of gene fusions exploiting miRNAs, J. Biomed. Inform. 129, 104057 (2022).
7. M. Lovino, M. S. Ciaburri, G. Urgese et al., DEEPrior: a deep learning tool for the prioritization of gene fusions, Bioinformatics 36, 3248 (2020).
8. S. Vincoff, S. Goel, K. Kholina et al., FusOn-pLM: a fusion oncoprotein-specific language model via adjusted rate masking, Nat. Commun. 16, 1436 (2025).
9. C.-C. Wu, H. C. Beird, J. Zhang and P. A. Futreal, FusionPathway: prediction of pathways and therapeutic targets associated with gene fusions in cancer, PLoS Comput. Biol. 14, e1006266 (2018).
10. K. Murakami, S.-I. Tago, S. Takishita et al., Pathogenicity prediction of gene fusion in structural variations: a knowledge graph-infused explainable artificial intelligence (XAI) framework, Cancers 16, 1915 (2024).
11. D. Cho, H.-M. Lee, J.-A. Kim et al., Autoinhibited protein database (AiPD): a curated database of autoinhibitory domains and their autoinhibition mechanisms, Database (Oxford) 2024, baae085 (2024).
12. S. Medves and J. B. Demoulin, Tyrosine kinase gene fusions in cancer: translating mechanisms into targeted therapies, J. Cell. Mol. Med. 16, 237 (2012).
13. E. De Braekeleer, N. Douet-Guilbert, D. Rowe et al., ABL1 fusion genes in hematological malignancies: a review, Eur. J. Haematol. 86, 361 (2011).
14. Z. Sondka, N. B. Dhir, D. Carvalho-Silva et al., COSMIC: a curated database of somatic variants and clinical data for cancer, Nucleic Acids Res. 52, D1210 (2024).
15. P. Flicek, M. R. Amode, D. Barrell et al., Ensembl 2014, Nucleic Acids Res. 42, D749 (2014).
16. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Res. 51, D523 (2023).
17. C. J. Clopper and E. S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26, 404 (1934).
18. V. Vovk, A. Gammerman and G. Shafer, Algorithmic Learning in a Random World, 2nd edn. (Springer, Cham, 2022).
19. A. N. Angelopoulos and S. Bates, A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification (2021).
20. D. Chakravarty, J. Gao, S. M. Phillips et al., OncoKB: a precision oncology knowledge base, JCO Precis. Oncol. 2017, PO.17.00011 (2017).
21. M. Griffith, N. C. Spies, K. Krysiak et al., CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer, Nat. Genet. 49, 170 (2017).

22. P. Horak, M. Griffith, A. M. Danos et al., Standards for the classification of pathogenicity of somatic variants in cancer (oncogenicity): joint recommendations of the Clinical Genome Resource (ClinGen), Cancer Genomics Consortium (CGC) and Variant Interpretation for Cancer Consortium (VICC), Genet. Med. 24, 986 (2022).
23. F. Weinberg, R. Griffin, M. Fröhlich et al., Identification and characterization of a BRAF fusion oncoprotein with retained autoinhibitory domains, Oncogene 39, 814 (2020).
24. A. N. Van, M. T. Kunkel, T. R. Baffi et al., Protein kinase C fusion proteins are paradoxically loss of function in cancer, J. Biol. Chem. 296, 100445 (2021).
25. K. M. Smith, R. Yacobi and R. A. Van Etten, Autoinhibition of Bcr-Abl through its SH3 domain, Mol. Cell 12, 27 (2003).
26. B. Nagar, O. Hantschel, M. A. Young et al., Structural basis for the autoinhibition of c-Abl tyrosine kinase, Cell 112, 859 (2003).
27. M. W. Richards, E. W. Law, L. P. Rennalls et al., Crystal structure of EML1 reveals the basis for Hsp90 dependence of oncogenic EML4-ALK by disruption of an atypical beta-propeller domain, Proc. Natl. Acad. Sci. U.S.A. 111, 5195 (2014).
28. J. A. Bridge, M. Kanamori, Z. Ma et al., Fusion of the ALK gene to the clathrin heavy chain gene, CLTC, in inflammatory myofibroblastic tumor, Am. J. Pathol. 159, 411 (2001).
29. P. Argani, M. Y. Lui, J. Couturier et al., A novel CLTC-TFE3 gene fusion in pediatric renal adenocarcinoma with t(X;17)(p11.2;q23), Oncogene 22, 5374 (2003).
30. C. Mayr, M. T. Hemann and D. P. Bartel, Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation, Science 315, 1576 (2007).
31. C. G. Woo, S. Seo, S. W. Kim et al., Differential protein stability and clinical responses of EML4-ALK fusion variants to various ALK inhibitors in advanced ALK-rearranged non-small cell lung cancer, Ann. Oncol. 28, 791 (2017).
32. M. W. Richards, L. O'Regan, D. Roth et al., Microtubule association of EML proteins and the EML4-ALK variant 3 oncoprotein require an N-terminal trimerization domain, Biochem. J. 467, 529 (2015).
33. H. Kurani, M. Gurav, O. Shetty et al., Pilocytic astrocytomas: BRAF V600E and BRAF fusion expression patterns in pediatric and adult age groups, Childs Nerv. Syst. 35, 1525 (2019).
34. C. Faulkner, H. P. Ellis, A. Shaw et al., BRAF fusion analysis in pilocytic astrocytomas: KIAA1549-BRAF 15-9 fusions are more frequent in the midline than within the cerebellum, J. Neuropathol. Exp. Neurol. 74, 867 (2015).
35. S. Barthelmess, H. Geddert, C. Boltze et al., Solitary fibrous tumors/hemangiopericytomas with different variants of the NAB2-STAT6 gene fusion are characterized by specific histomorphology and distinct clinicopathological features, Am. J. Pathol. 184, 1209 (2014).
36. K. Akaike, A. Kurisaki-Arakawa, K. Hara et al., Distinct clinicopathological features of NAB2-STAT6 fusion gene variants in solitary fibrous tumor with emphasis on the acquisition of highly malignant potential, Hum. Pathol. 46, 347 (2015).
37. S. V. Tavtigian, M. S. Greenblatt, S. M. Harrison et al., Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework, Genet. Med. 20, 1054 (2018).