Bio-curation for cellular signalling

Russ Harmer (CNRS, Lyon)

Who am I?

- Mathematician / computer scientist
  - formal semantics & graph-based knowledge representation
    - theory of knowledge update
    - generic methods for graph databases
    - my favourite use case: cellular signalling

Cellular signalling

- Largely de-centralized co-ordination of tissue micro-architecture
  - *inter*-cellular signals and *intra*-cellular transduction
  - build a body: morphogenesis
  - maintain a body: morphostasis — and its disruption

Intra-cellular signalling

very briefly

- Signal transduction
  - membrane-associated receptors capture signals
  - conformational change induces enzymatic activity
  - cascade of PTMs and assembly of protein complexes
  - modulation of gene expression

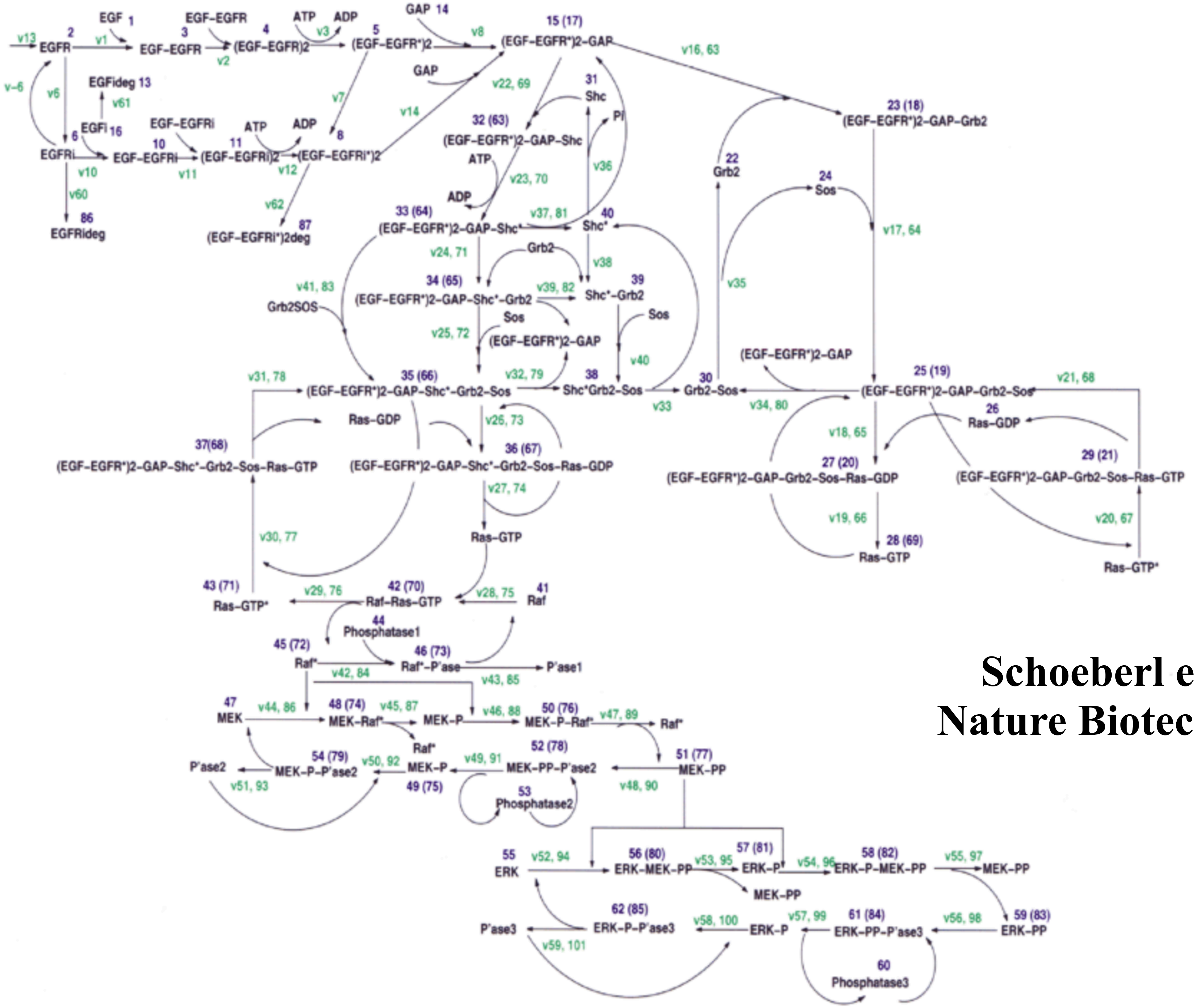
Intra-cellular signalling

modelling issues

- How does a perturbation affect signalling?
  - over/under-expression, mutations
  - with respect to a notion of behaviour: trajectory/pathway
- Effect of perturbation must not be hard-wired
  - transitions defined at the level of protoforms
  - only mention the known necessary conditions

Reaction-based

manual Reading, manual Assembly



Schoeberl et al. (2002),
Nature Biotechnology 20(4)

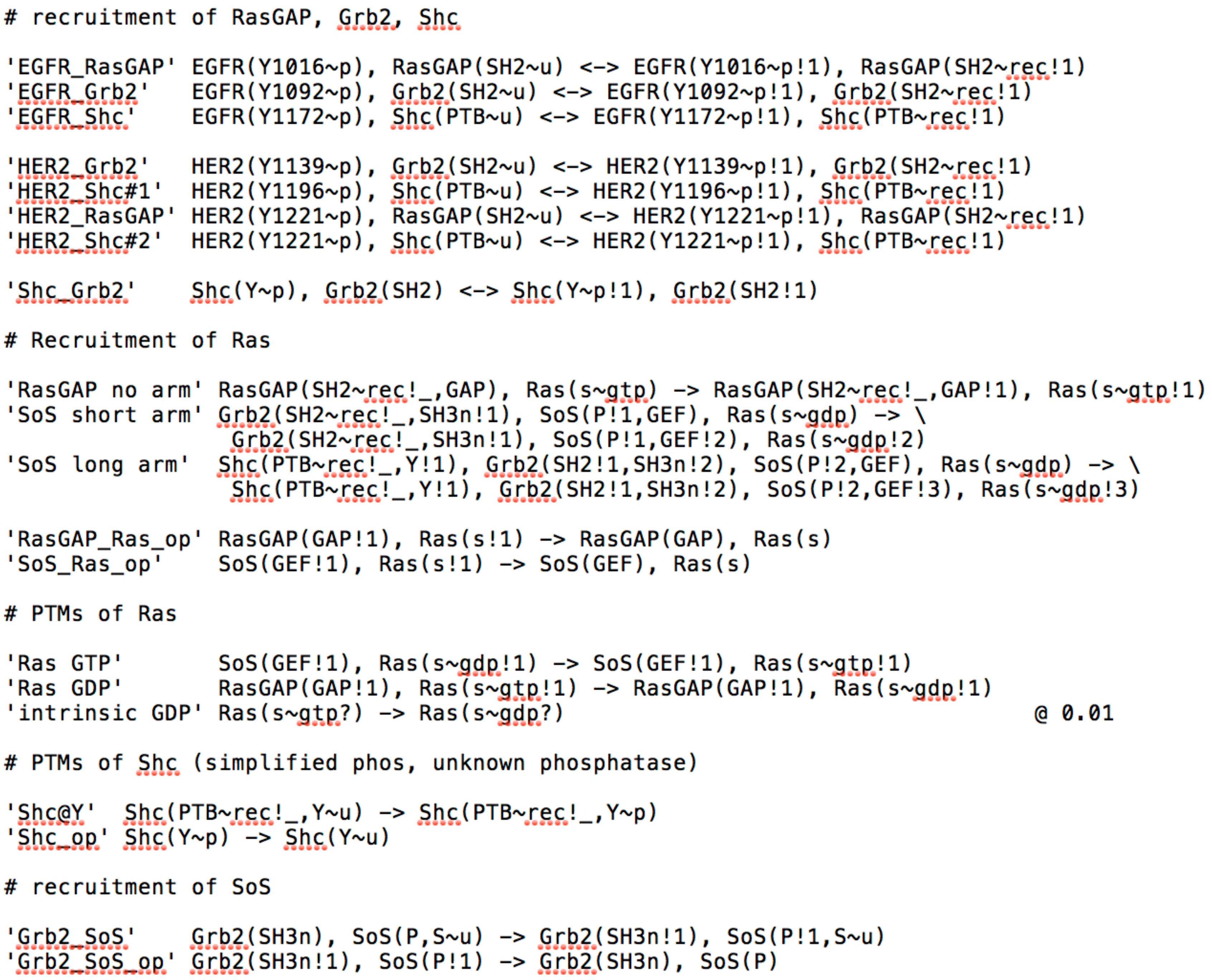
Intra-cellular signalling

rule-based modelling

- Rule-based modelling partially enables this
  - agents have sites (nodes) that can carry edges and/or states
  - transitions are instances of graph rewriting rules
  - pathways as causal traces
- Cell type-dependence
  - the same signal provokes different results in different cells
  - what is a cell type anyway? is it a helpful concept?

Rule-based modelling

manual Reading, manual Assembly



```
# recruitment of RasGAP, Grb2, Shc

'EGFR_RasGAP' EGFR(Y1016~p), RasGAP(SH2~u) <-> EGFR(Y1016~p!1), RasGAP(SH2~rec!1)
'EGFR_Grb2'   EGFR(Y1092~p), Grb2(SH2~u) <-> EGFR(Y1092~p!1), Grb2(SH2~rec!1)
'EGFR_Shc'    EGFR(Y1172~p), Shc(PTB~u) <-> EGFR(Y1172~p!1), Shc(PTB~rec!1)

'HER2_Grb2'   HER2(Y1139~p), Grb2(SH2~u) <-> HER2(Y1139~p!1), Grb2(SH2~rec!1)
'HER2_Shc#1'  HER2(Y1196~p), Shc(PTB~u) <-> HER2(Y1196~p!1), Shc(PTB~rec!1)
'HER2_RasGAP' HER2(Y1221~p), RasGAP(SH2~u) <-> HER2(Y1221~p!1), RasGAP(SH2~rec!1)
'HER2_Shc#2'  HER2(Y1221~p), Shc(PTB~u) <-> HER2(Y1221~p!1), Shc(PTB~rec!1)

'Shc_Grb2'    Shc(Y~p), Grb2(SH2) <-> Shc(Y~p!1), Grb2(SH2!1)

# Recruitment of Ras

'RasGAP no arm' RasGAP(SH2~rec!_,GAP), Ras(s~gtp) -> RasGAP(SH2~rec!_,GAP!1), Ras(s~gtp!1)
'SoS short arm' Grb2(SH2~rec!_,SH3n!1), SoS(P!1,GEF), Ras(s~gdp) -> \
                 Grb2(SH2~rec!_,SH3n!1), SoS(P!1,GEF!2), Ras(s~gdp!2)
'SoS long arm'  Shc(PTB~rec!_,Y!1), Grb2(SH2!1,SH3n!2), SoS(P!2,GEF), Ras(s~gdp) -> \
                 Shc(PTB~rec!_,Y!1), Grb2(SH2!1,SH3n!2), SoS(P!2,GEF!3), Ras(s~gdp!3)

'RasGAP_Ras_op' RasGAP(GAP!1), Ras(s!1) -> RasGAP(GAP), Ras(s)
'SoS_Ras_op'    SoS(GEF!1), Ras(s!1) -> SoS(GEF), Ras(s)

# PTMs of Ras

'Ras GTP'       SoS(GEF!1), Ras(s~gdp!1) -> SoS(GEF!1), Ras(s~gtp!1)
'Ras GDP'       RasGAP(GAP!1), Ras(s~gtp!1) -> RasGAP(GAP!1), Ras(s~gdp!1)
'intrinsic GDP' Ras(s~gtp?) -> Ras(s~gdp?)                                   @ 0.01

# PTMs of Shc (simplified phos, unknown phosphatase)

'Shc@Y'  Shc(PTB~rec!_,Y~u) -> Shc(PTB~rec!_,Y~p)
'Shc_op' Shc(Y~p) -> Shc(Y~u)

# recruitment of SoS

'Grb2_SoS'    Grb2(SH3n), SoS(P,S~u) -> Grb2(SH3n!1), SoS(P!1,S~u)
'Grb2_SoS_op' Grb2(SH3n!1), SoS(P!1) -> Grb2(SH3n), SoS(P)
```

Intra-cellular signalling

knowledge deficit

- We don’t know all the details of signalling PPIs
  - and what we do know is scattered across the literature
- The knowledge we do have is trapped
  - in PubMed — with a citation bias
  - in the heads of biologists — with a selection bias

Bio-curation

reclaiming knowledge

- Aggregate necessary conditions for PPIs…
  - find a paper and extract knowledge about a PPI
  - does this update our knowledge of that PPI?
- …in de-contextualized fashion
  - at the level of the ensemble of products of a gene
  - determine automatically which PPIs for which proteins

Bio-curation

reclaiming knowledge

- Huge cognitive and logistic burden for a curator
  - time-consuming (and biased) to find and read the papers
  - need to extract knowledge, keep track of updates, …
- To what end?
  - don’t just want to feed a database
  - we want to animate this knowledge…

Bio-curation

executable knowledge

- Instantiation of our knowledge (at time *t*)
  - choice of proteins: splice variants, mutants, …
  - automatic generation of a rule-based model
- Investigation of its consequences
  - which pathways does this signal activate (or not)?
  - selection bias test — does this agree with the biologists?

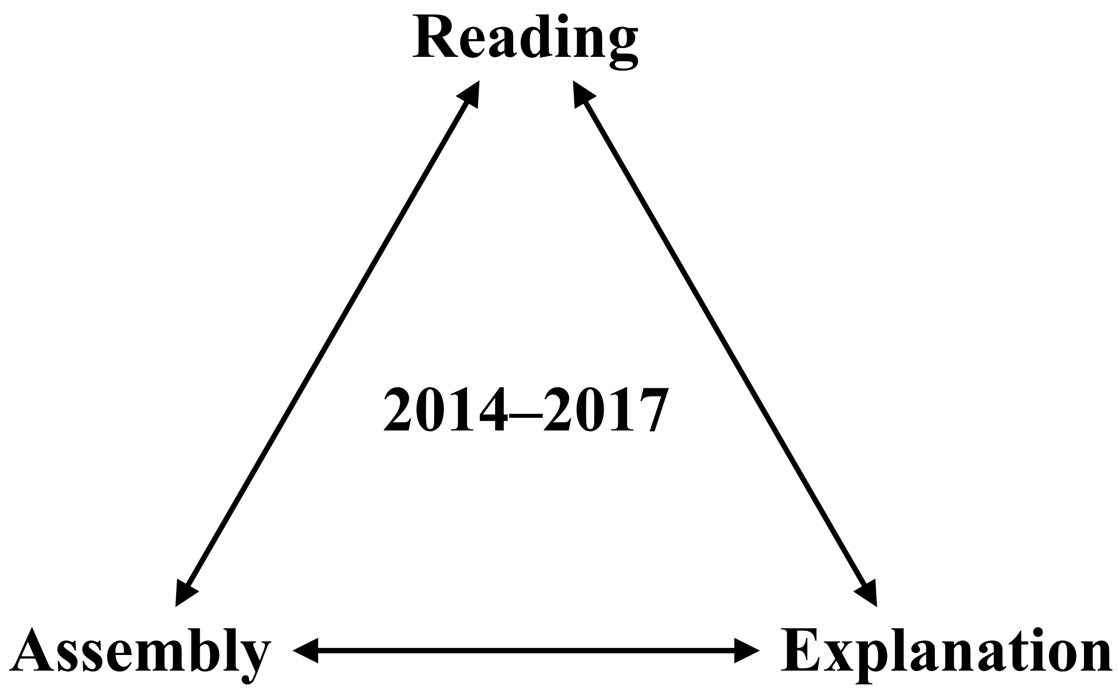
Bio-curation

executable knowledge

- Modelling as a tool for discovery
  - a model is not an artifact that codifies understanding
  - but rather a permanent work in progress that seeks understanding
- The process of modelling, not the model…

Big Mechanism

DARPA



Reading

2014–2017

Assembly

Explanation

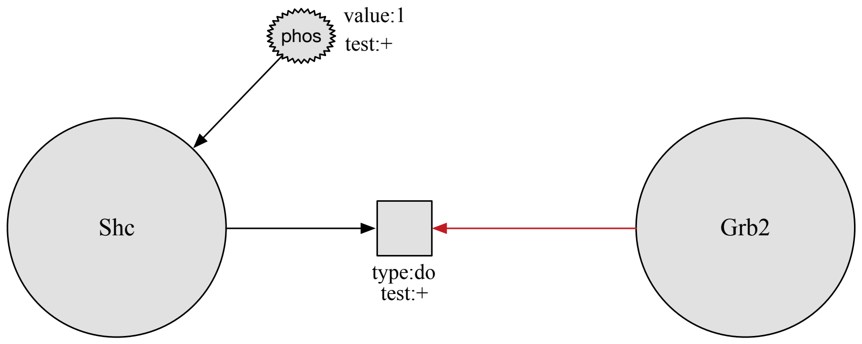
KAMI

semi-automatic Assembly

- Graph-based knowledge representation
  - knowledge update and aggregation via graph rewriting
  - instantiation of knowledge via protein definitions (also via graph rewriting)
- Automatic generation of a rule-based Kappa model
  - to reconstruct pathways from simulations

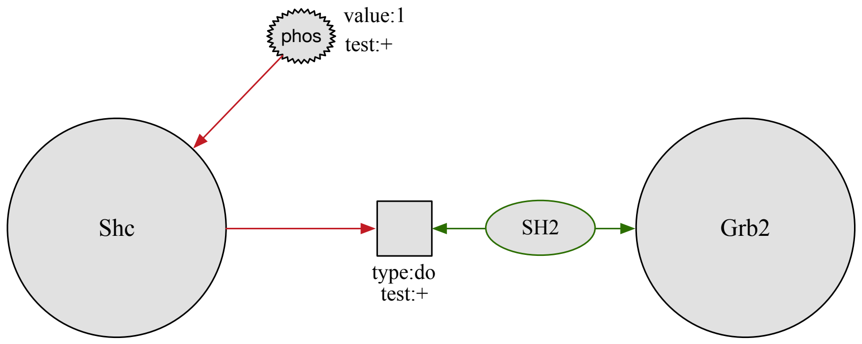
Knowledge update

“phosphorylated Shc binds Grb2”



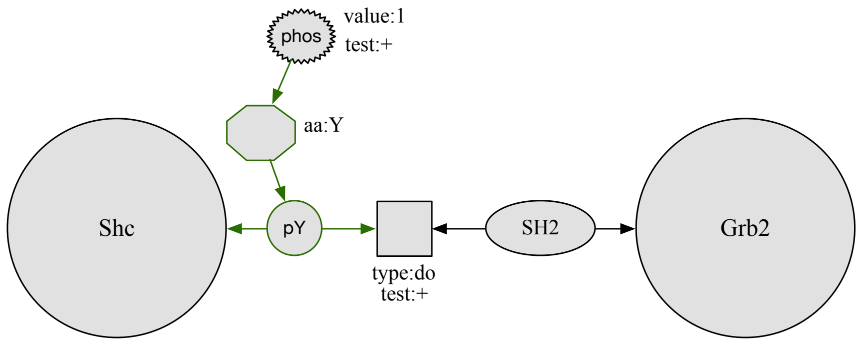
manual update
by the curator

“phosphorylated Shc binds
the SH2 domain of Grb2”

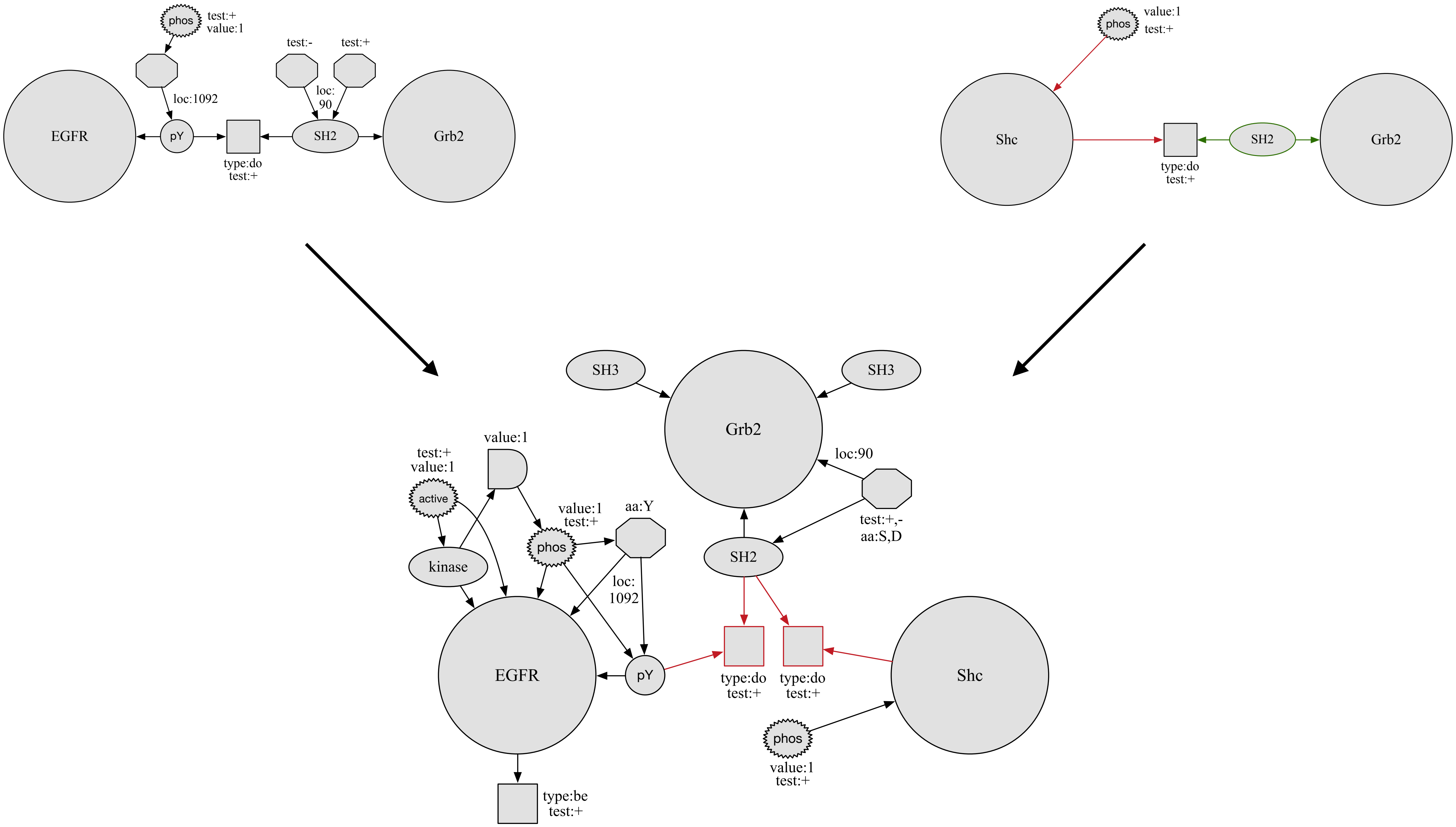


automatic update
from background knowledge

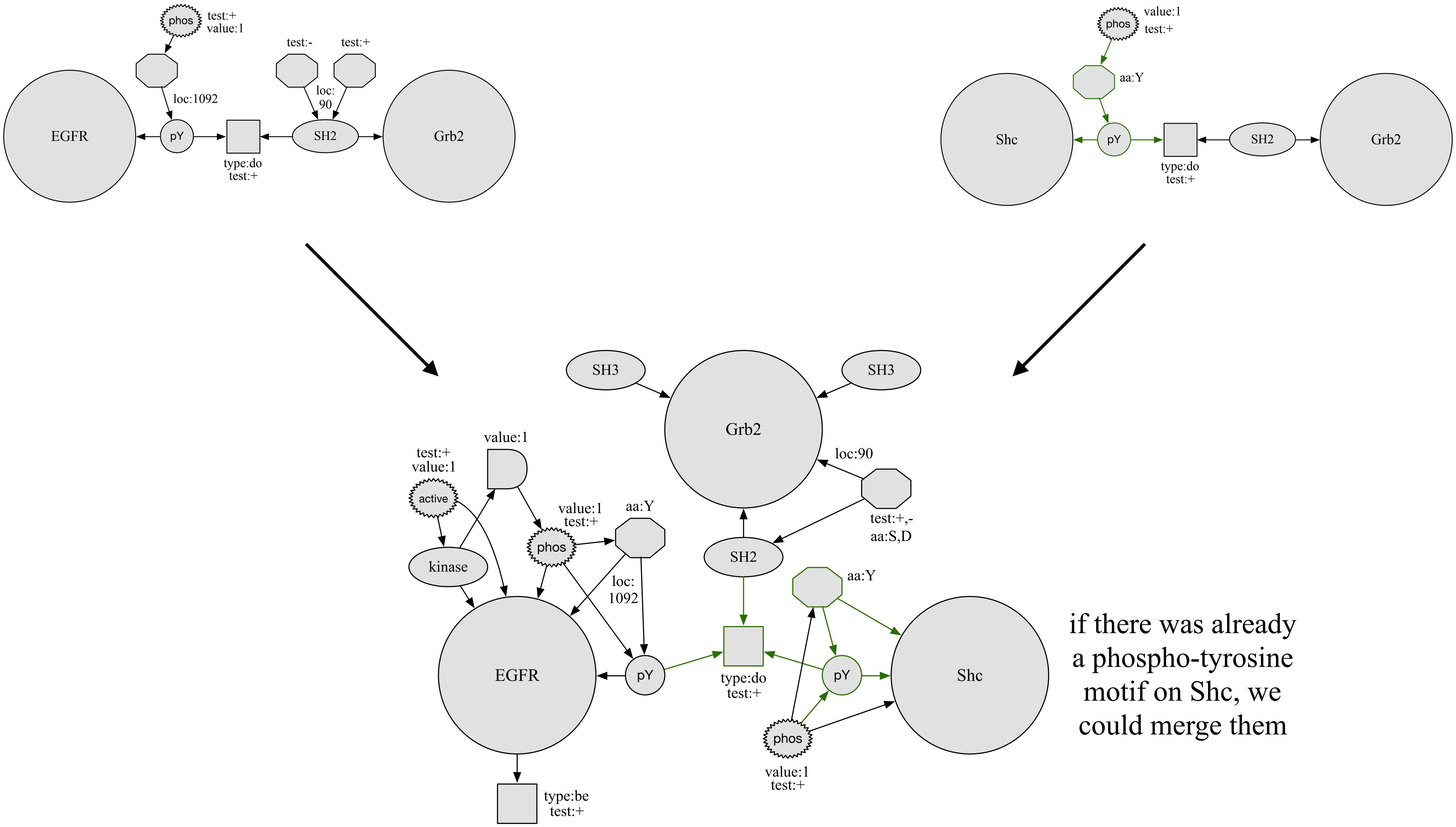
“a phospho-tyrosine motif of Shc
binds the SH2 domain of Grb2”



Knowledge Aggregation

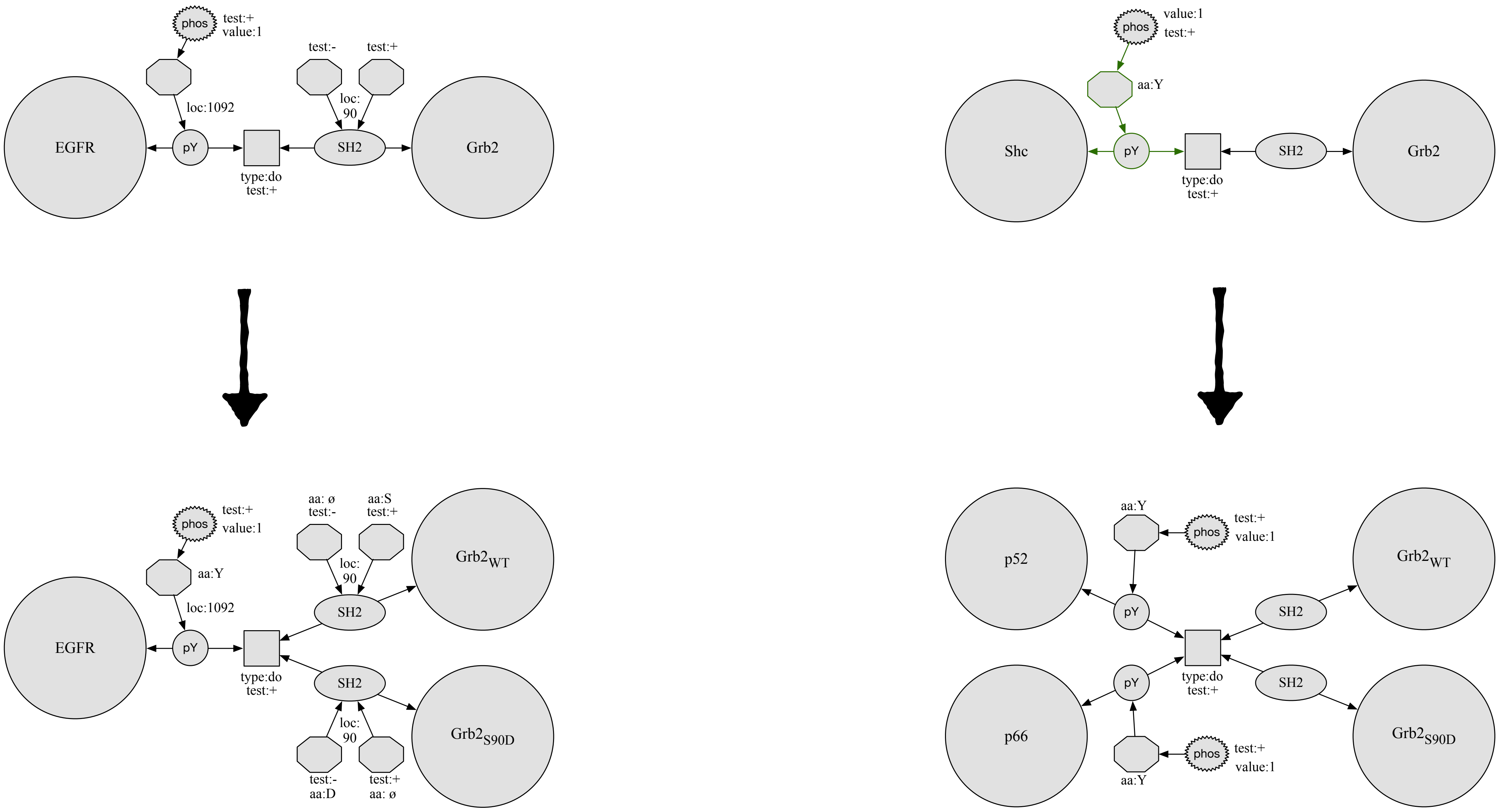


Knowledge Aggregation

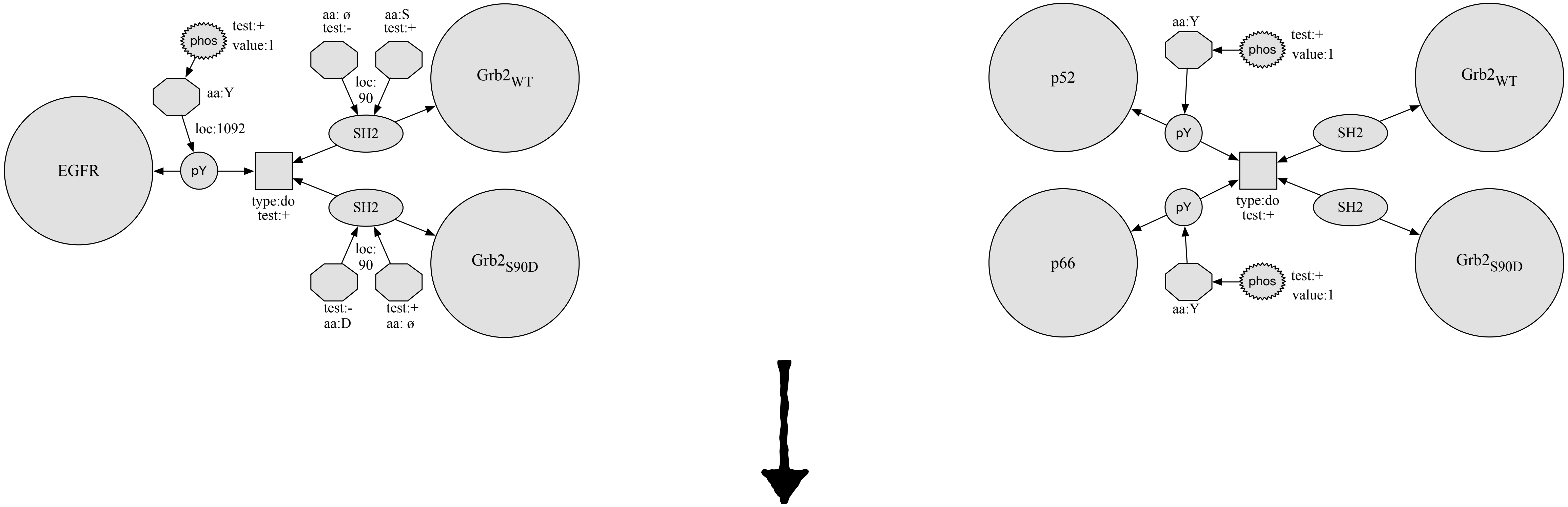


if there was already
a phospho-tyrosine
motif on Shc, we
could merge them

Model Instantiation



Model Instantiation



```
p52(pY,Y~1), Grb2WT(sh2) ->
 p52(pY!0,Y~1), Grb2WT(sh2!0)

p66(pY,Y~1), Grb2WT(sh2) ->
 p66(pY!0,Y~1), Grb2WT(sh2!0)

p52(pY,Y~1), Grb2S90D(sh2) ->
 p52(pY!0,Y~1), Grb2S90D(sh2!0)

p66(pY,Y~1), Grb2S90D(sh2) ->
 p66(pY!0,Y~1), Grb2S90D(sh2!0)

EGFR(pY,Y1092~1), Grb2WT(sh2) ->
 EGFR(pY,Y1092~1), Grb2WT(sh2!0)

# no rule for Grb2S90D
```

this is Kappa code

KAMI

semi-automatic Assembly

- Knowledge update and aggregation
    - meaningful updates exploiting background knowledge
    - one action per mechanism: grounding for PPIs
- Model instantiation
    - many models from a single knowledge corpus

KAMI++

current work

- Reclaiming further knowledge
  - combine knowledge from multiple species: region homology and by similarity inference
  - updates acquire an epistemic status: observed vs. inferred
- Greater representational power
  - kinetic refinements, phenomenological definitions and assertions

Conclusions

- Why build complicated models?
  - one body of knowledge that profitably instantiates to several contexts >> multiple independent curation efforts
  - hard to build useful simple models — instead try to simplify in a specific context in a principled manner
- Why seek simple models?
  - clarify what is really important — what is a cell type?
  - ultimately to address inter-cellular signalling