EnnCore

MANCHESTER
1824
The University of Manchester

Explainability & Inference Controls

Andre Freitas
ExplAIn Lab

idiap
RESEARCH INSTITUTE

Julia Rozanova

Marco Valentino

Edoardo Manino

Lucas Cordeiro

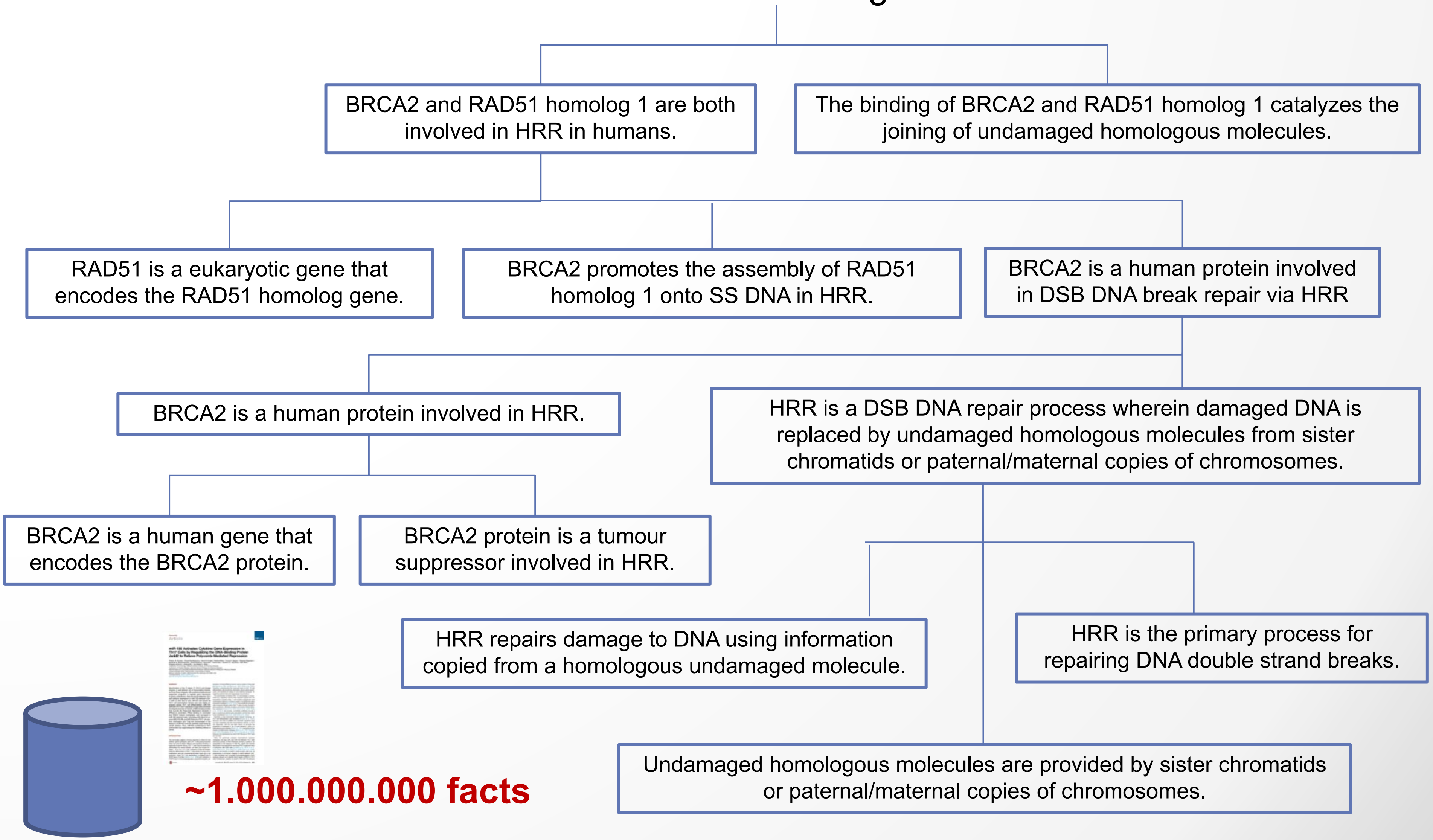
Danilo Carvalho

Giangiacomo Mercatali

Mokanarangan Thayaparan

Expert-level scientific inference & explanation

Claim: BRCA2 promotes the joining of undamaged homologous repair molecules via RAD51 homolog 1 in humans.



BRCA2 and RAD51 homolog 1 are both involved in HRR in humans.

The binding of BRCA2 and RAD51 homolog 1 catalyzes the joining of undamaged homologous molecules.

RAD51 is a eukaryotic gene that encodes the RAD51 homolog gene.

BRCA2 promotes the assembly of RAD51 homolog 1 onto SS DNA in HRR.

BRCA2 is a human protein involved in DSB DNA break repair via HRR

BRCA2 is a human protein involved in HRR.

HRR is a DSB DNA repair process wherein damaged DNA is replaced by undamaged homologous molecules from sister chromatids or paternal/maternal copies of chromosomes.

BRCA2 is a human gene that encodes the BRCA2 protein.

BRCA2 protein is a tumour suppressor involved in HRR.

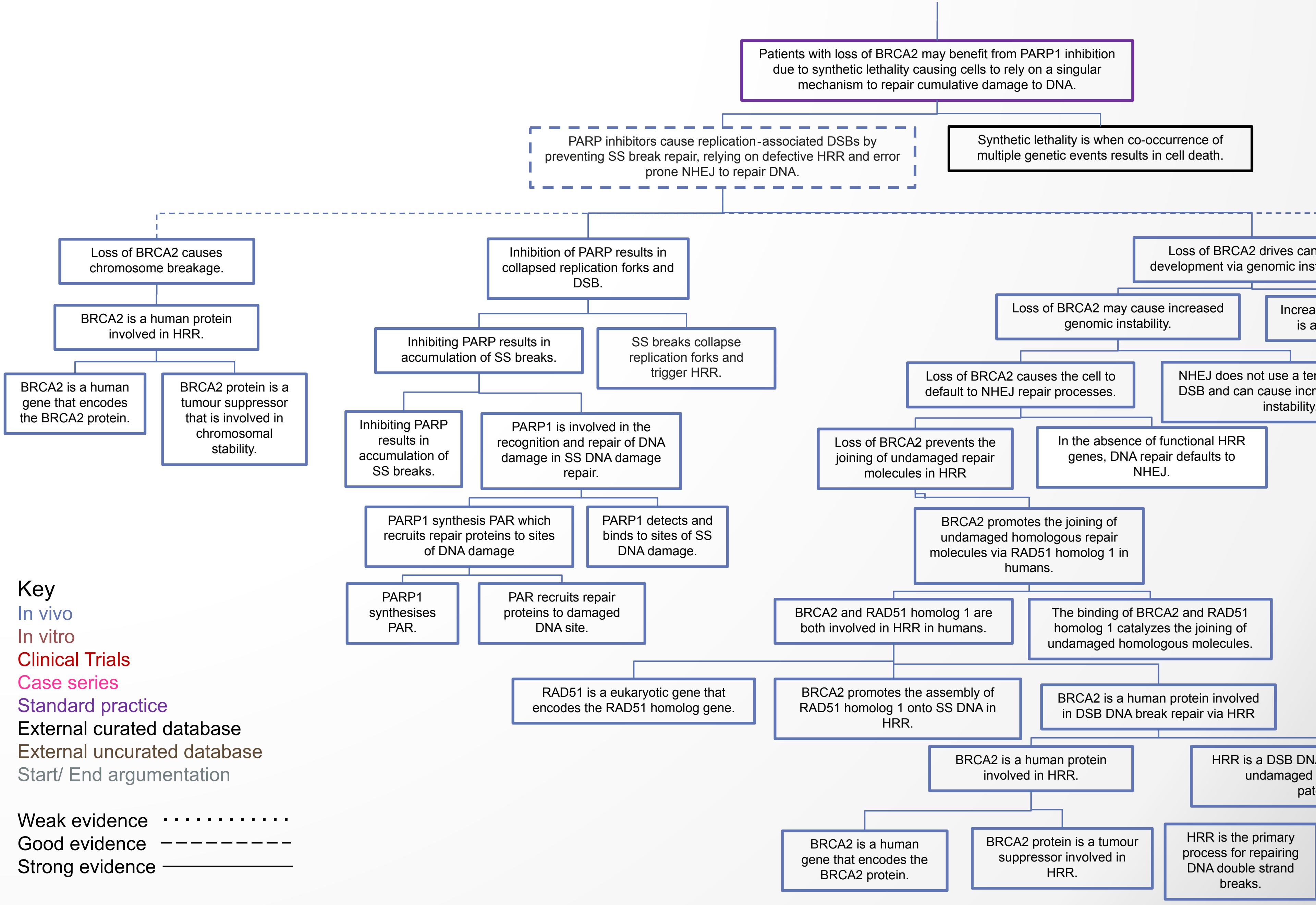
HRR repairs damage to DNA using information copied from a homologous undamaged molecule.

HRR is the primary process for repairing DNA double strand breaks.

Undamaged homologous molecules are provided by sister chromatids or paternal/maternal copies of chromosomes.

~1.000.000.000 facts

Prostate cancer patient with loss of BRCA2 may benefit from PARP1 inhibition



Patients with loss of BRCA2 may benefit from PARP1 inhibition due to synthetic lethality causing cells to rely on a singular mechanism to repair cumulative damage to DNA.

PARP inhibitors cause replication-associated DSBs by preventing SS break repair, relying on defective HRR and error prone NHEJ to repair DNA.

Synthetic lethality is when co-occurrence of multiple genetic events results in cell death.

Loss of BRCA2 causes chromosome breakage.

BRCA2 is a human protein involved in HRR.

BRCA2 is a human gene that encodes the BRCA2 protein.

BRCA2 protein is a tumour suppressor that is involved in chromosomal stability.

Inhibition of PARP results in collapsed replication forks and DSB.

Inhibiting PARP results in accumulation of SS breaks.

SS breaks collapse replication forks and trigger HRR.

Inhibiting PARP results in accumulation of SS breaks.

PARP1 is involved in the recognition and repair of DNA damage in SS DNA damage repair.

PARP1 synthesis PAR which recruits repair proteins to sites of DNA damage

PARP1 detects and binds to sites of SS DNA damage.

PARP1 synthesises PAR.

PAR recruits repair proteins to damaged DNA site.

Loss of BRCA2 drives can
development via genomic ins

Loss of BRCA2 may cause increased genomic instability.

Increa
is a

Loss of BRCA2 causes the cell to default to NHEJ repair processes.

NHEJ does not use a te
DSB and can cause incr
instability

Loss of BRCA2 prevents the joining of undamaged repair molecules in HRR

In the absence of functional HRR genes, DNA repair defaults to NHEJ.

BRCA2 promotes the joining of undamaged homologous repair molecules via RAD51 homolog 1 in humans.

BRCA2 and RAD51 homolog 1 are both involved in HRR in humans.

The binding of BRCA2 and RAD51 homolog 1 catalyzes the joining of undamaged homologous molecules.

RAD51 is a eukaryotic gene that encodes the RAD51 homolog gene.

BRCA2 promotes the assembly of RAD51 homolog 1 onto SS DNA in HRR.

BRCA2 is a human protein involved in DSB DNA break repair via HRR

BRCA2 is a human protein involved in HRR.

HRR is a DSB DN
undamaged
pat

BRCA2 is a human gene that encodes the BRCA2 protein.

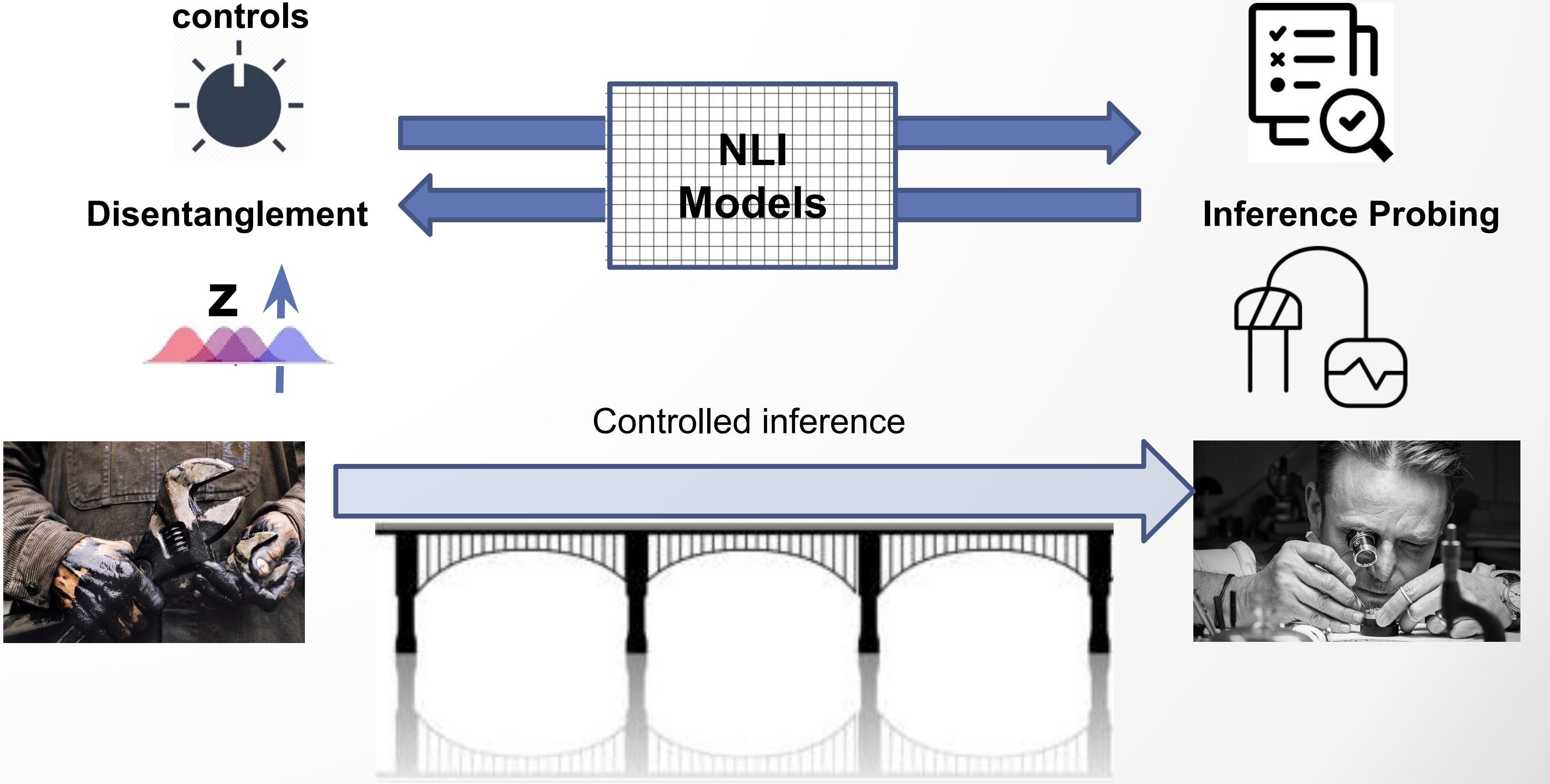
BRCA2 protein is a tumour suppressor involved in HRR.

HRR is the primary process for repairing DNA double strand breaks.

Key
In vivo
In vitro
Clinical Trials
Case series
Standard practice
External curated database
External uncurated database
Start/ End argumentation

Weak evidence · · · · · · · · · · · ·
Good evidence – – – – – – – – –
Strong evidence ————————

Controlled Inference



Intervention

Encoding inference controls

Disentanglement

Observation

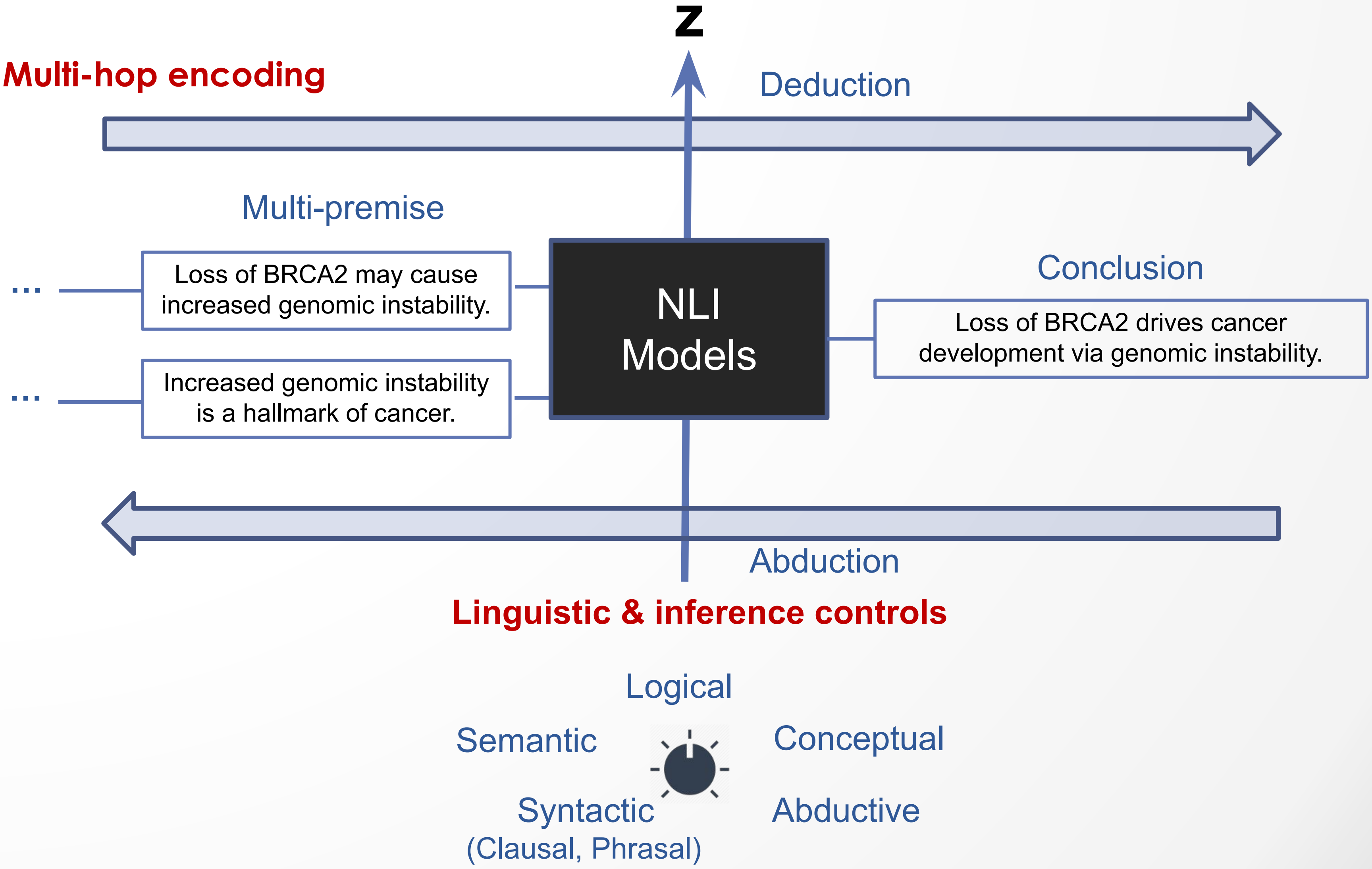
Metamorphic Testing

Inference Probing

Controlled inference

4 strategic pillars

Encoding Inference Controls



z

Multi-hop encoding

Deduction

Multi-premise

...

Loss of BRCA2 may cause increased genomic instability.

...

Increased genomic instability is a hallmark of cancer.

NLI Models

Conclusion

Loss of BRCA2 drives cancer development via genomic instability.

Abduction

Linguistic & inference controls

Logical

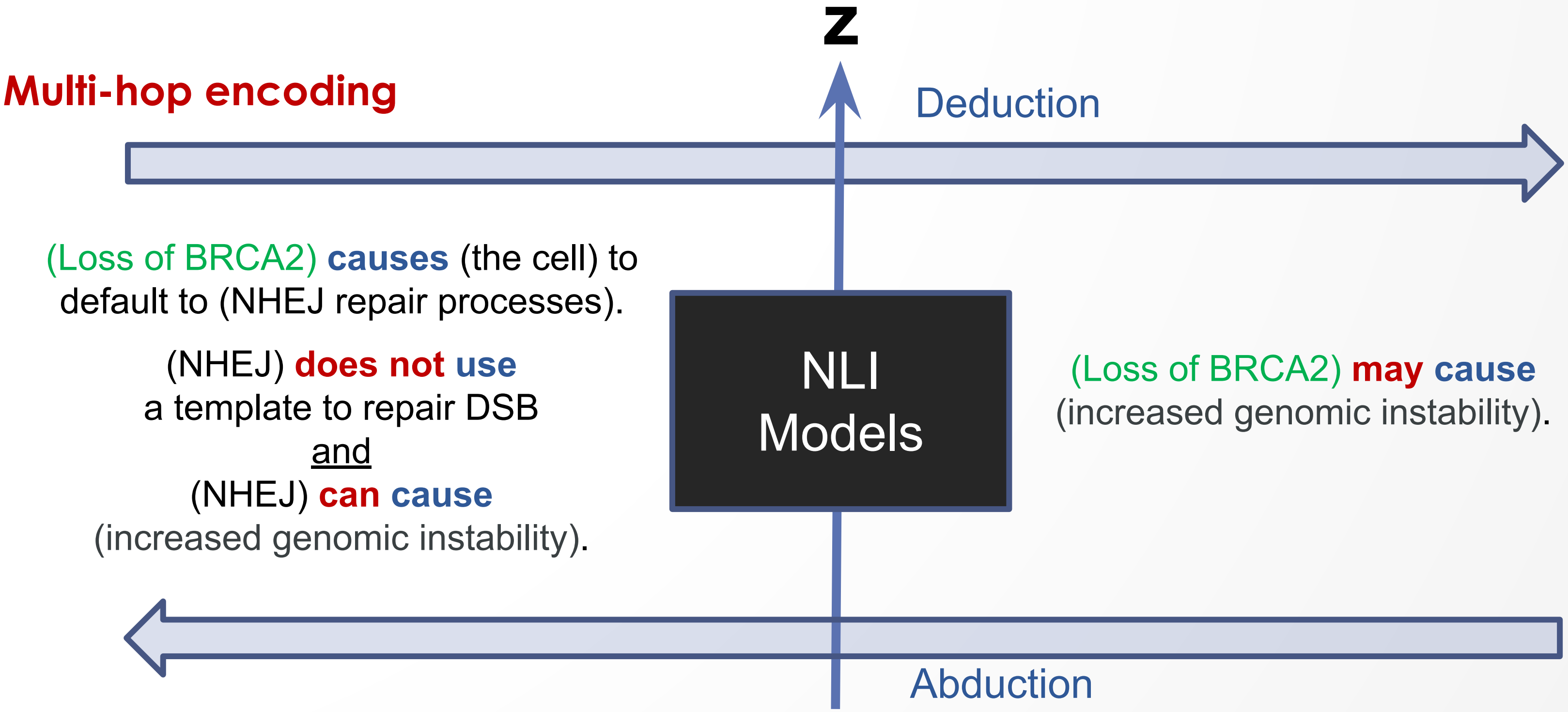
Semantic

Conceptual

Syntactic (Clausal, Phrasal)

Abductive

Encoding Inference Controls



z

Multi-hop encoding

Deduction

(Loss of BRCA2) **causes** (the cell) to default to (NHEJ repair processes).

(NHEJ) **does not use** a template to repair DSB
and
(NHEJ) **can cause** (increased genomic instability).

NLI Models

(Loss of BRCA2) **may cause** (increased genomic instability).

Abduction

Linguistic & inference controls

Logical

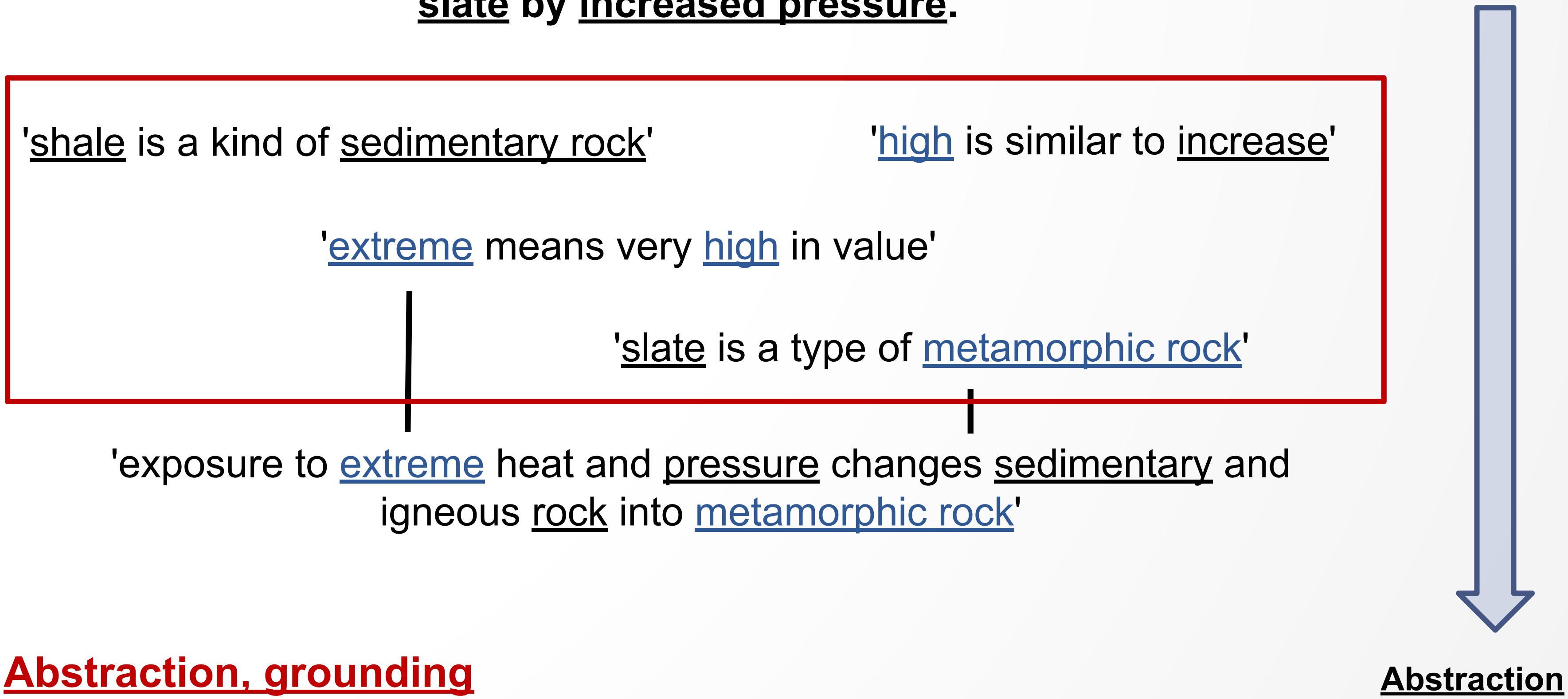
Semantic

Conceptual

Syntactic
(Clausal, Phrasal)

Abductive

H: Shale is a sedimentary rock that can be metamorphosed into slate by increased pressure.



'shale is a kind of sedimentary rock'

'high is similar to increase'

'extreme means very high in value'

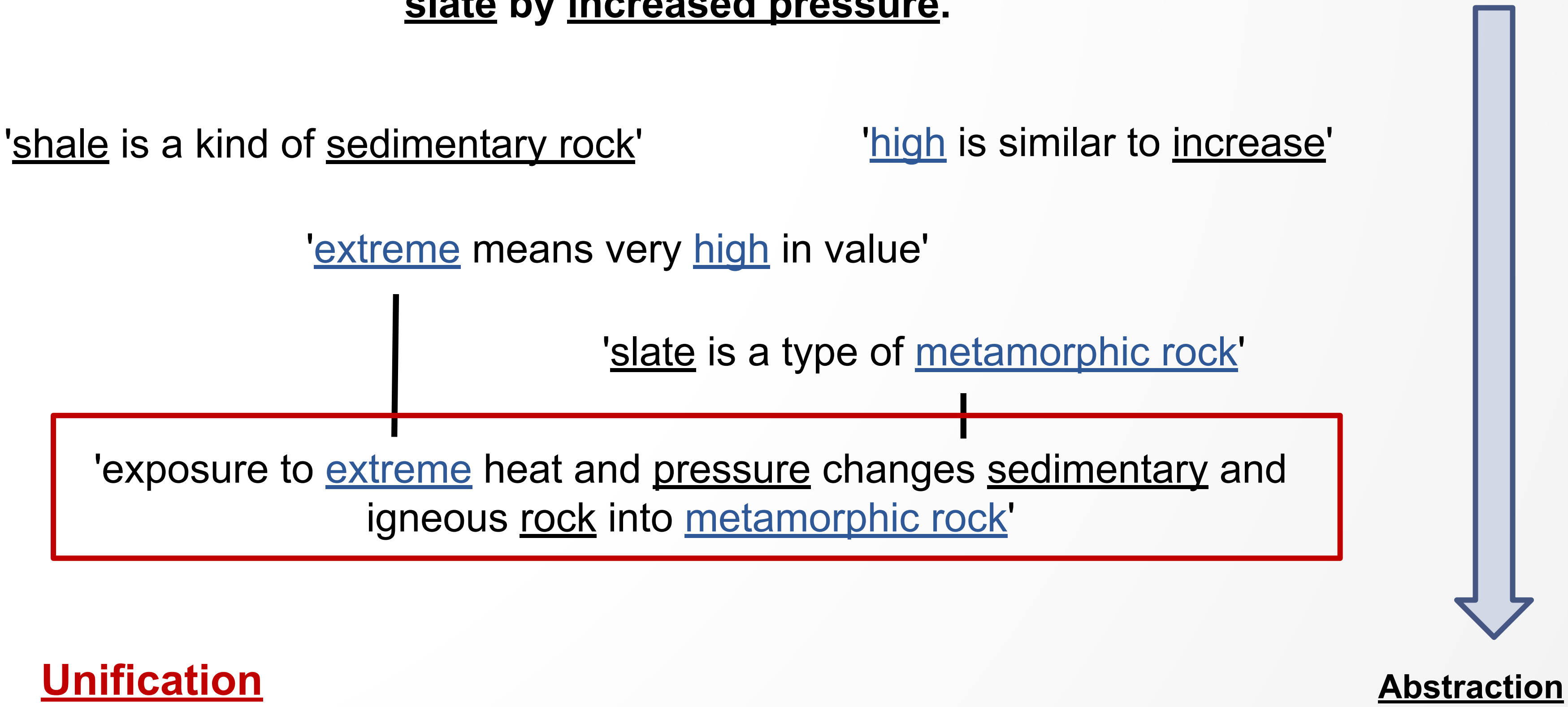
'slate is a type of metamorphic rock'

'exposure to extreme heat and pressure changes sedimentary and igneous rock into metamorphic rock'

Abstraction

Abstraction, grounding

H: Shale is a sedimentary rock that can be metamorphosed into slate by increased pressure.



'shale is a kind of sedimentary rock'

'high is similar to increase'

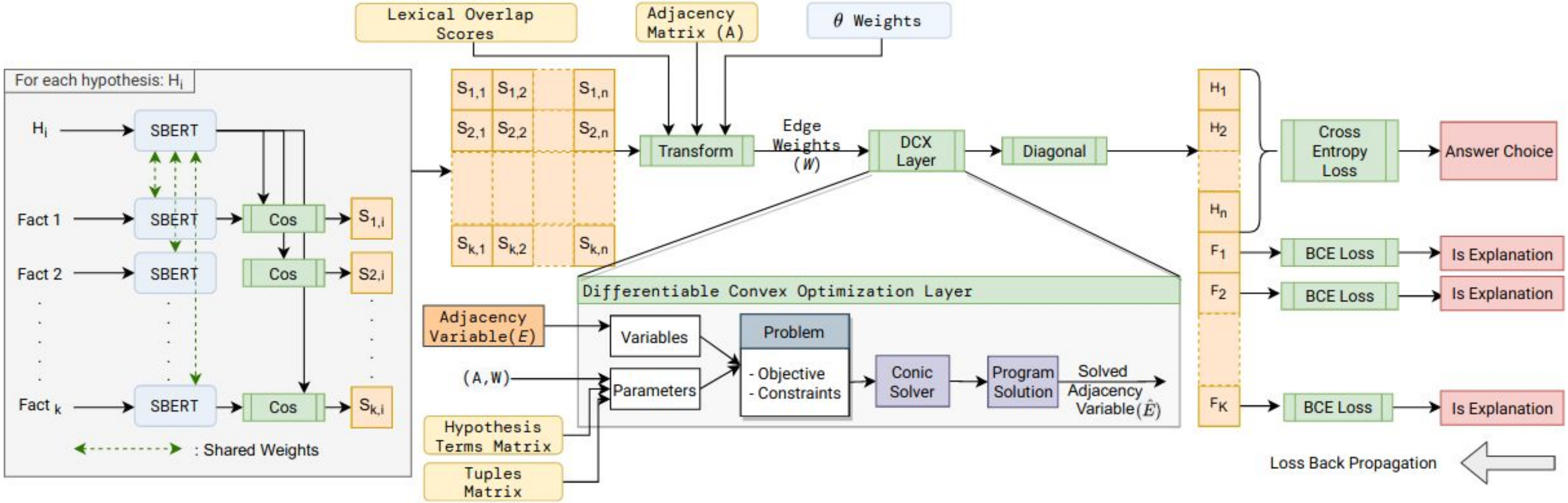
'extreme means very high in value'

'slate is a type of metamorphic rock'

'exposure to extreme heat and pressure changes sedimentary and igneous rock into metamorphic rock'

Unification

Abstraction



An end-to-end differentiable framework that incorporates constraints via convex optimization layers into broader transformers-based architectures.

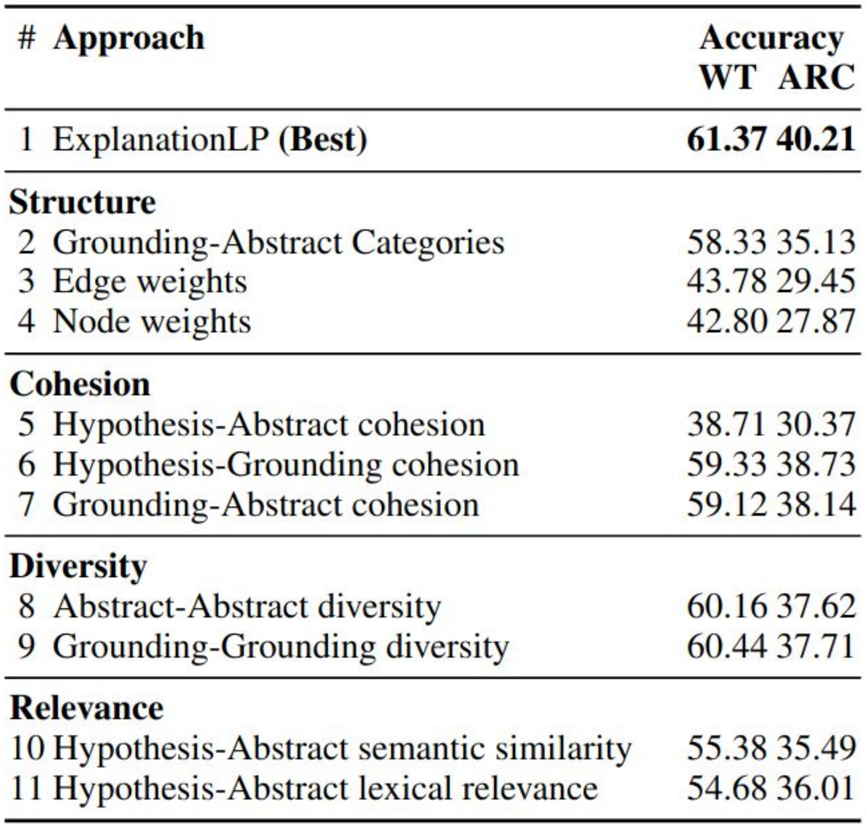
Direction of a programmable abductive NLI Solver

Explainable Inference Over Grounding-Abstract Chains for Science Questions

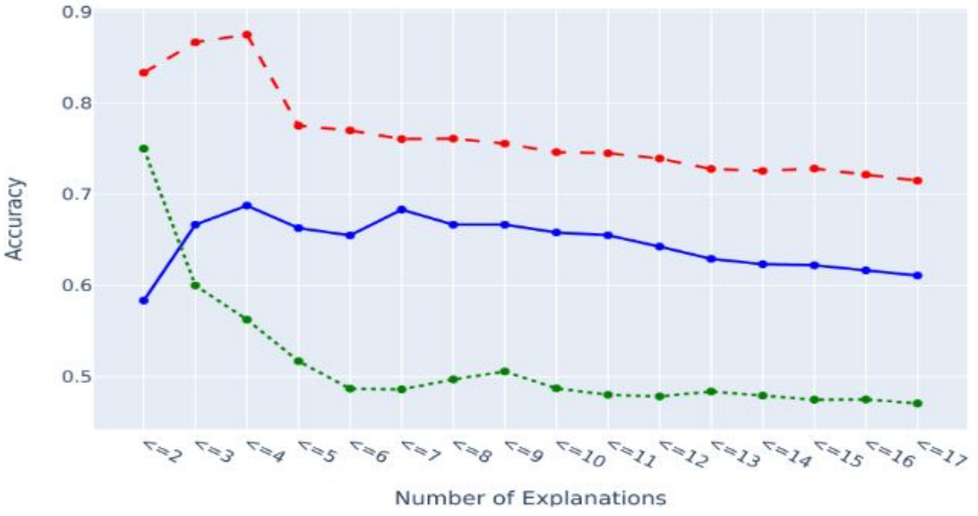
Thayaparan et al., ACL Findings (2021)

∂-Explainer: Abductive Natural Language Inference via Differentiable Convex Optimization

Thayaparan et al., ArXiv 2105.03417 (2021)



#	Approach	Accuracy	
		WT	ARC
1	ExplanationLP (**Best**)	**61.37**	**40.21**
	Structure		
2	Grounding-Abstract Categories	58.33	35.13
3	Edge weights	43.78	29.45
4	Node weights	42.80	27.87
	Cohesion		
5	Hypothesis-Abstract cohesion	38.71	30.37
6	Hypothesis-Grounding cohesion	59.33	38.73
7	Grounding-Abstract cohesion	59.12	38.14
	Diversity		
8	Abstract-Abstract diversity	60.16	37.62
9	Grounding-Grounding diversity	60.44	37.71
	Relevance		
10	Hypothesis-Abstract semantic similarity	55.38	35.49
11	Hypothesis-Abstract lexical relevance	54.68	36.01



red: ExplanationLP + UR

blue: BERT$_{Large}$ + UR

green: PathNet + UR

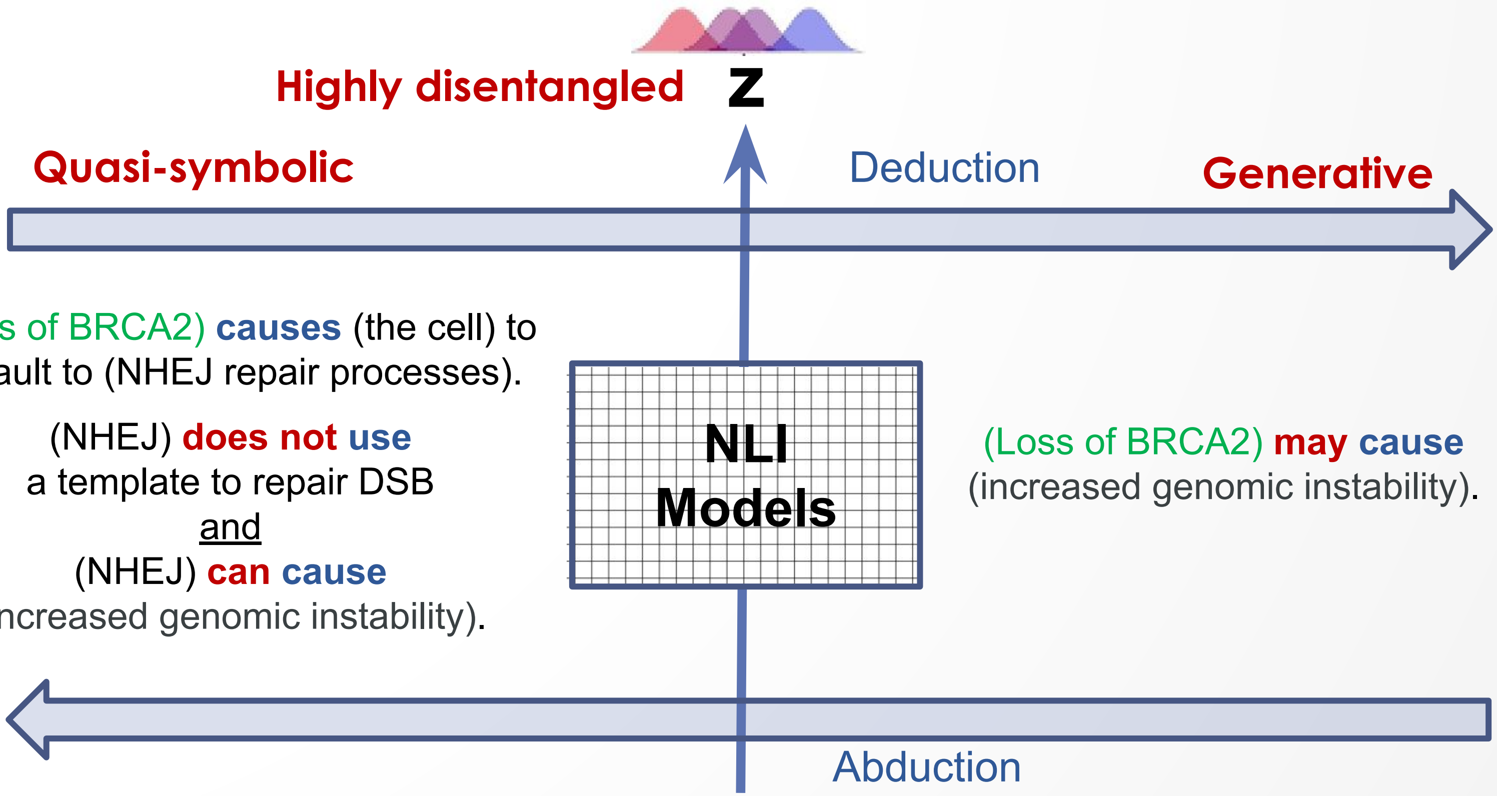
# of parameters:

- BERTBase: 110M parameters
- BERTLarge: 340M parameters
- ExplanationLP: 9 parameters

Explainable Inference Over Grounding-Abstract Chains for Science Questions

Thayaparan et al., ACL Findings (2021)

Disentanglement



Highly disentangled **z**

Quasi-symbolic

Deduction

Generative

(Loss of BRCA2) **causes** (the cell) to default to (NHEJ repair processes).

(NHEJ) **does not use** a template to repair DSB

and

(NHEJ) **can cause** (increased genomic instability).

NLI Models

(Loss of BRCA2) **may cause** (increased genomic instability).

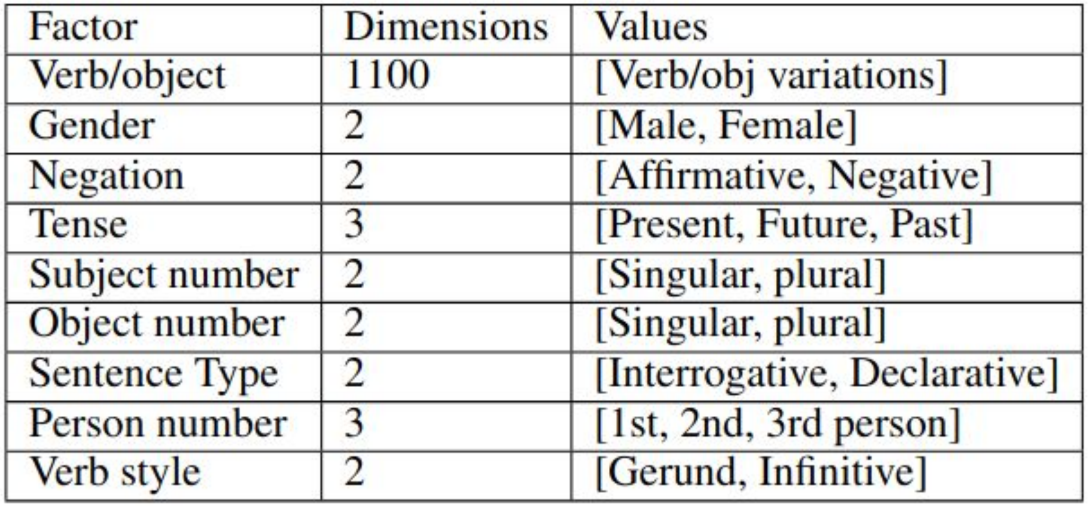
Abduction

Disentangling Generative Factors in Natural Language with Discrete Variational Autoencoders

Mercatalli & Freitas, EMNLP Findings (2021)

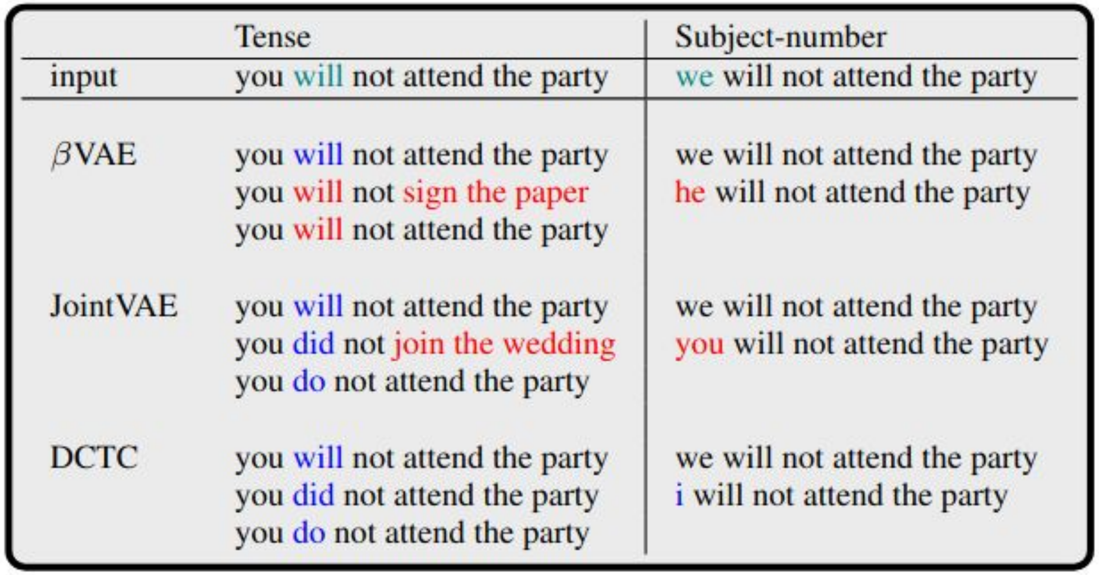
Disentangling Generative Factors in Natural Language with Discrete Variational Autoencoders

Mercatalli & Freitas, EMNLP Findings (2021)



Factor	Dimensions	Values
Verb/object	1100	[Verb/obj variations]
Gender	2	[Male, Female]
Negation	2	[Affirmative, Negative]
Tense	3	[Present, Future, Past]
Subject number	2	[Singular, plural]
Object number	2	[Singular, plural]
Sentence Type	2	[Interrogative, Declarative]
Person number	3	[1st, 2nd, 3rd person]
Verb style	2	[Gerund, Infinitive]

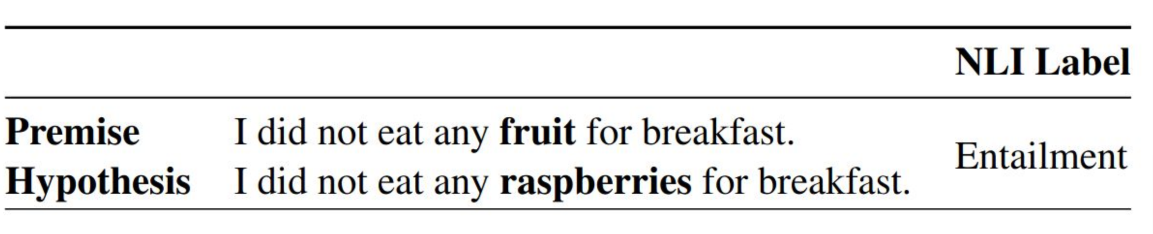
Latent traversal



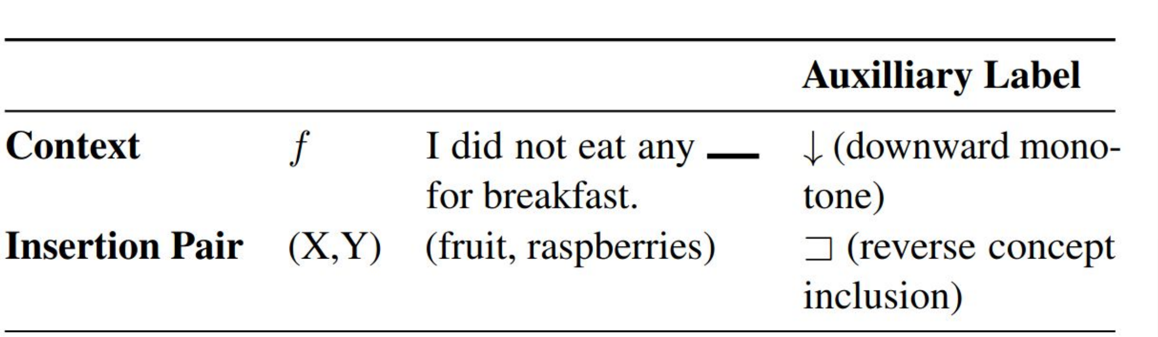
	Tense	Subject-number
input	you will not attend the party	we will not attend the party
βVAE	you will not attend the party you will not sign the paper you will not attend the party	we will not attend the party he will not attend the party
JointVAE	you will not attend the party you did not join the wedding you do not attend the party	we will not attend the party you will not attend the party
DCTC	you will not attend the party you did not attend the party you do not attend the party	we will not attend the party i will not attend the party

Inference Probing

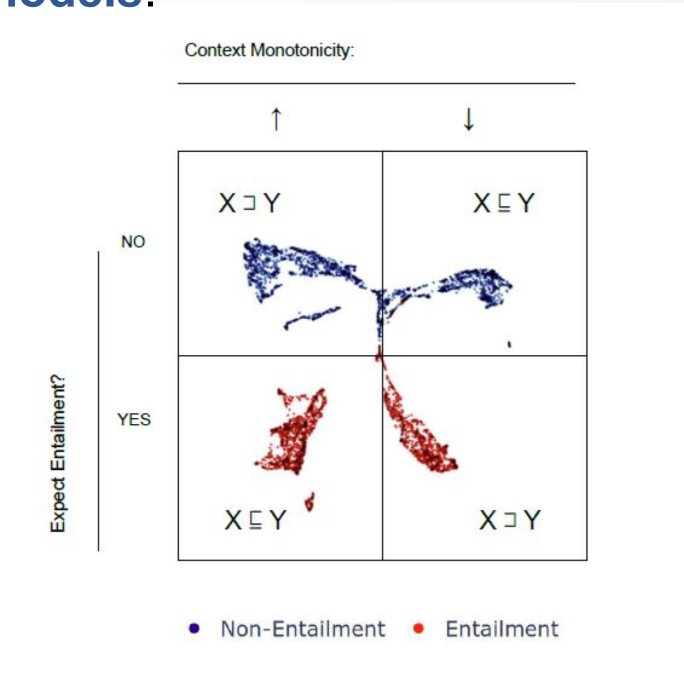
Structural investigation as to whether the behaviour of natural logic formalisms are mimicked within popular **transformer-based NLI models**.



		NLI Label
Premise	I did not eat any **fruit** for breakfast.	Entailment
Hypothesis	I did not eat any **raspberries** for breakfast.	



			Auxilliary Label
Context	*f*	I did not eat any ___ for breakfast.	↓ (downward monotone)
Insertion Pair	(X,Y)	(fruit, raspberries)	⊐ (reverse concept inclusion)

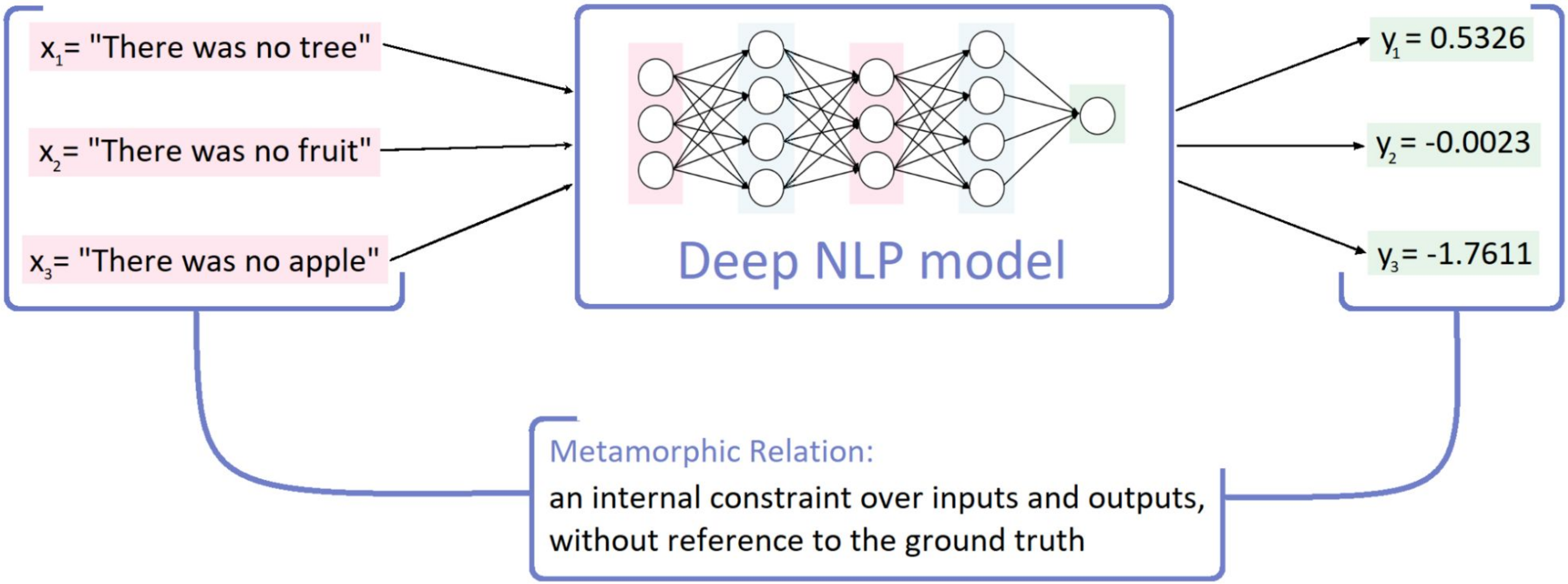


Well-known NLI models demonstrate a systematic failure to model context monotonicity, but they can be fine-tuned to integrate this behaviour.

Decomposing Natural Logic Inferences in Neural NLI Rozanova et al., (2021)

Does My Representation Capture X? Probe-Ably Ferreira et al., ACL Demo (2021)

Metamorphic Testing



x_1= "There was no tree"
x_2= "There was no fruit"
x_3= "There was no apple"
Deep NLP model
y_1 = 0.5326
y_2 = -0.0023
y_3 = -1.7611
Metamorphic Relation:
an internal constraint over inputs and outputs, without reference to the ground truth

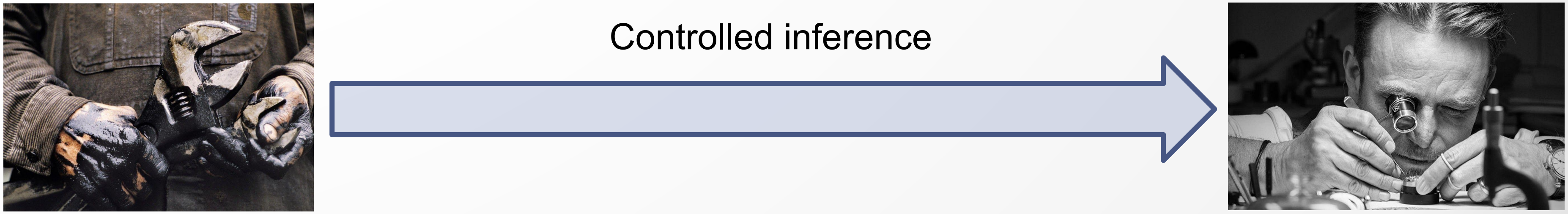
Systematicity, Compositionality and Transitivity of Deep NLP Models: a Metamorphic Testing Perspective,

Manino et al., ACL Findings (2022)

Take away

Explainable, controlled, neuro-symbolic inference

- Exploiting the structure of abstract inference for multi-hop inference design.
- Declarative solvers: encoding strategies for complex and abstract inference.
- Disentanglement: interpretability and quasi-symbolic behavior.
- Model behaviour: inference probing and metamorphic testing.


Controlled inference