Explainability & Inference Controls

Andre Freitas
ExplAIIn Lab
**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.

- 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.

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

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

- BRCA2 is a human protein involved in HRR.

- BRCA2 protein is a tumour suppressor involved in HRR.

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

**~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.

BRCA2 is a human gene that encodes the BRCA2 protein.

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

BRCA2 is a human protein involved in HRR.

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

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

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

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.

PARP1 detects and binds to sites of SS DNA damage.

PARP1 synthesises PAR which recruits repair proteins to sites of DNA damage.

PARP1 synthesises PAR which recruits repair proteins to sites of DNA damage.

In vitro

In vivo

Clinical Trials

Case series

Standard practice

External curated database

External uncurated database

Start/ End argumentation

Weak evidence

Good evidence

Strong evidence

Inhibiting PARP results in accumulation of SS breaks.

Inhibiting PARP results in accumulation of SS breaks.

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

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

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.

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

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

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

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.

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

BRCA2 protein is a tumour suppressor involved in HRR.

BRCA2 is a human gene that encodes the BRCA2 protein.

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

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

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

Loss of BRCA2 drives cancer development via genomic instability.

Loss of BRCA2 may cause increased genomic instability.

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

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

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

BRCA2 is a human protein that encodes the BRCA2 protein.

BRCA2 is a human gene that encodes the BRCA2 protein.

BRCA2 is a human protein involved in HRR.

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

Inhibiting PARP results in accumulation of SS breaks.

Inhibiting PARP results in accumulation of SS breaks.

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

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

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.

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

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

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

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.

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

BRCA2 protein is a tumour suppressor involved in HRR.

BRCA2 is a human gene that encodes the BRCA2 protein.

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

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

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

Loss of BRCA2 drives cancer development via genomic instability.

Loss of BRCA2 may cause increased genomic instability.

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

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

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

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.

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

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

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

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

Loss of BRCA2 drives cancer development via genomic instability.

Loss of BRCA2 may cause increased genomic instability.

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

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

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

Encoding inference controls

Disentanglement

Observation

Metamorphic Testing

Inference Probing

NLI Models

Controlled inference

4 strategic pillars
Loss of BRCA2 drives cancer development via genomic instability.

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.

Linguistic & inference controls

- Logical
- Semantic
- Syntactic (Clausal, Phrasal)
- Conceptual
- Abductive
(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.

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

*α-Explainer: Abductive Natural Language Inference via Differentiable Convex Optimization*

Thayaparan et al., ArXiv 2105.03417 (2021)
<table>
<thead>
<tr>
<th>#</th>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WT  ARC</td>
</tr>
<tr>
<td>1</td>
<td>ExplanationLP (Best)</td>
<td>61.37 40.21</td>
</tr>
<tr>
<td>2</td>
<td>Grounding-Abstract Categories</td>
<td>58.33 35.13</td>
</tr>
<tr>
<td>3</td>
<td>Edge weights</td>
<td>43.78 29.45</td>
</tr>
<tr>
<td>4</td>
<td>Node weights</td>
<td>42.80 27.87</td>
</tr>
<tr>
<td>5</td>
<td>Hypothesis-Abstract cohesion</td>
<td>38.71 30.37</td>
</tr>
<tr>
<td>6</td>
<td>Hypothesis-Grounding cohesion</td>
<td>59.33 38.73</td>
</tr>
<tr>
<td>7</td>
<td>Grounding-Abstract cohesion</td>
<td>59.12 38.14</td>
</tr>
<tr>
<td>8</td>
<td>Abstract-Abstract diversity</td>
<td>60.16 37.62</td>
</tr>
<tr>
<td>9</td>
<td>Grounding-Grounding diversity</td>
<td>60.44 37.71</td>
</tr>
<tr>
<td>10</td>
<td>Hypothesis-Abstract semantic similarity</td>
<td>55.38 35.49</td>
</tr>
<tr>
<td>11</td>
<td>Hypothesis-Abstract lexical relevance</td>
<td>54.68 36.01</td>
</tr>
</tbody>
</table>

# 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)
(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).

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

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

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

**Latent traversal**

<table>
<thead>
<tr>
<th>Tense</th>
<th>Subject-number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>input</strong></td>
<td>you will not attend the party</td>
</tr>
<tr>
<td>(\beta\text{VAE})</td>
<td>you will not attend the party</td>
</tr>
<tr>
<td></td>
<td>you will not sign the paper</td>
</tr>
<tr>
<td>Joint(\beta\text{VAE})</td>
<td>you will not attend the party</td>
</tr>
<tr>
<td></td>
<td>you did not join the wedding</td>
</tr>
<tr>
<td></td>
<td>you do not attend the party</td>
</tr>
<tr>
<td>DCTC</td>
<td>you will not attend the party</td>
</tr>
<tr>
<td></td>
<td>you did not attend the party</td>
</tr>
<tr>
<td></td>
<td>you do not attend the party</td>
</tr>
</tbody>
</table>
Inference Probing

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

<table>
<thead>
<tr>
<th>Premise</th>
<th>I did not eat any <strong>fruit</strong> for breakfast.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypothesis</td>
<td>I did not eat any <strong>raspberries</strong> for breakfast.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Context</th>
<th>$f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I did not eat any ___ for breakfast.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Insertion Pair</th>
<th>(X,Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(fruit, raspberries)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Auxilliary Label</th>
<th>$\downarrow$ (downward monotone)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Expect Entailment?</th>
<th>NO</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>X &amp; Y</td>
<td>X &amp; Y</td>
<td></td>
</tr>
</tbody>
</table>

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

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