

Sony AI「Hakken」が報告した老化関連の未知遺伝子関係の深掘り調査

Executive Summary

Sony AIの研究チームが2026年9月3日にarXivへ公開した「Hakken: Predicting future discoveries to fill the gaps in today's knowledge」は、科学文献から時系列知識グラフを構築し、グラフニューラルネットワークとLLMを組み合わせて「まだ文献上に明示されていない関係」を予測するシステムを報告した。生物医学への実証では、老化関連遺伝子として専門家が選定した1,386エンティティから959,805の未順序ペアを作り、信頼度閾値を超える約154万件の関係仮説を生成した。その後、フィルタリングと専門家選定を経て3仮説をヒト細胞株で実験し、**TP53-BAMBI**と**RAF1-TNF**の2関係を著者らは「Supported」と判定、**SOAT1-STAT3**は否定された。¹

ただし、ニュース見出しから受ける「AIが老化遺伝子を新発見し、その老化機能を証明した」という理解は強すぎる。今回実験されたのは、**老化関連遺伝子の候補集合内で予測された機能的な発現関係**であり、細胞老化、寿命、健康寿命、老化時計、senescence markerなどの老化表現型そのものは測定されていない。実験はHepG2およびTHP-1というヒト細胞株で行われ、老齢個体・若齢個体比較、一次細胞、動物モデル、臨床検体による検証は報告されていない。したがって現時点で科学的に最も妥当な表現は、「**AIが老化関連の探索空間から、従来Hakkenの知識グラフに記録されていなかった二つの機能的遺伝子関係を予測し、限定的な細胞実験で支持する結果を得た**」である。²

さらに「2/3が成功したから66.7%の発見成功率」と解釈することもできない。約154万候補から既知関係を除き、各エンティティ上位3件に絞って2,804件とし、その後**生物医学専門家が信頼度0.8超の候補から3件を主観的に選んだ**ためである。この3件はランダム標本ではなく、非常に強い選択バイアスを持つ。³

実験エビデンスにも重要な温度差がある。TP53-BAMBIは、MDM2阻害剤Nutlin-3によるp53誘導に伴いBAMBI mRNAが高濃度・長時間条件で増え、再実験でも再現されたため「機能的関連」は比較的説得力がある。一方、TP53のCRISPR knockout、rescue、BAMBIプロモーターへのp53直接結合を示すChIP/CUT&RUNなどは行われておらず、直接転写制御までは証明されていない。RAF1-TNFはさらに複雑で、RAF1阻害剤GW5074ではTNF mRNAが増えたものの、別のpan-RAF阻害剤AZ628では一時的に逆方向へ変化し、TNF分泌タンパク質には変化がなく、RAF阻害剤による「paradoxical ERK activation」も観測された。よって、**RAF1がTNFを直接抑制するという正確な関係述語の証拠は「限定的～中程度」と評価すべきである**。⁴

Hakkenの技術的価値はむしろ、単純なLLMによるアイデア生成とは異なり、**時間を明示した知識グラフ、positive-unlabeled learning、LLM意味情報、関係タイプ予測、説明経路PHEInE、過去時点に戻した歴史的バックテスト**を一つのシステムに統合した点にある。2020年カットオフ評価ではTHiGERLLMはmacro-F1で既存手法を上回ったものの、すべての指標で優位だったわけではなく、前身のグラフ中心モデルTHiGERがweighted-F1やnDCGなどで勝る指標もある。この点からも「全面的なSOTA」と単純化するのは適切でない。⁵

また、2026年9月10日時点では発表から約1週間しか経っておらず、論文は**arXivプレプリントで未査読**である。Sony ResearchはHakkenOSSを公開しているが、本論文の主要な生物医学評価データにはPMC Open Access、MEDLINEに加え選択的なライセンスデータが含まれる一方、研究者向けGUIではライセンス上の理由からPubTator3のみを用いた版を提供したと論文が説明している。そのため、コード公開は大きな前進であるものの、**論文の主要結果を第三者が完全にbit-for-bit再現できる状態とは言いにくい**。⁶

調査項目別の結論

調査項目	調査結果	本レポート評価
発表の要旨	2026年9月3日、Sony AI等の研究者がarXiv公開。約154万の老化関連仮説から3件をwet-labへ、TP53-BAMBIとRAF1-TNFを支持と報告。 7	重要だがプレプリント段階
Hakken技術	Temporal KG + GNN/Temporal Transformer + Mistral-7B系 LLM + PU learning + PHELIxE説明。 8	技術的新規性は高い
未知関係	TP53→BAMBI「affects expression」、RAF1→TNF「decreases expression」。SOAT1→STAT3は不支持。 9	関係の存在と正確な因果方向を分けて評価すべき
一次データ・追試	論文内で追加確認実験あり。ただし独立大学等による外部追試・査読論文は現時点で確認できない。 4	内部再現あり／独立再現なし
専門家評価	GUIを7組織の研究者が評価し定性的には好評。ただしHakkenを名指しした独立査読・専門家批評はまだ確認できない。 10	外部評価はまだ早期
限界・バイアス	候補選択、文献出版バイアス、抽出誤差、LLM時系列リーク、薬理学的off-target、因果説明不可など。 11	重大、特に“発見率”解釈に注意
推奨アクション	遺伝学的perturbation、複数研究室再現、老化モデル、事前登録したprospective benchmarkが必要。	最優先
ソース	arXiv原論文・SonyResearch GitHubを最優先。Sony AI公式はプロジェクト背景を補完。日本語報道はAIDB/XenoSpectrum等が先行。 12	一次情報中心に読むべき

調査範囲と発表の位置づけ

何が、いつ、誰によって発表されたのか

中心となる一次資料は、Tarek R. Besold、Uchenna Akujuobi、Pablo Sanchez、Alessandra Toniato、Kana Maruyama、Jihun Choi、Samy Badreddineら計25名が著者となるarXiv論文 **“Hakken: Predicting future discoveries to fill the gaps in today's knowledge”** で、2026年9月3日に投稿された。著者所属はSony AIのBarcelona、Zurich、Tokyo、London、New York拠点に加え、Systems Biology InstituteやOISTなどを含む。論文はHakkenを「現在の知識の演繹的閉包を超え、まだ文献に記録されていない新関係を予測する domain-agnostic prediction and explanation system」と位置づける。 13

Sony AI公式のScientific Discoveryページは、同チームの目的を、データ・知識駆動型手法と研究者向けソフトウェアを用いて科学者のideationを支援し、科学的発見を加速することと説明している。2026年1月のSony AI公式ブログでも、科学文献を静的な文書集合ではなく「時間とともに変化する相互接続システム」として扱い、研究者の仮説評価を支援する方向性が説明されている。したがってHakkenは突然出現した単発研究というより、Sony AIが数年進めてきたliterature-based hypothesis generationの延長線上にある。 14

一方、調査時点のSony AI公式News／Publicationsページでは、Hakkenという名称の独立したプレスリリース・出版ページを確認できなかった。したがって、**Hakkenそのものについて最も正式かつ詳細な一次発表は現状arXiv論文とSonyResearch公式GitHubリポジトリと考えるのが妥当である**。Sony AI公式Webサイトは研究プログラムの背景資料として扱うべきである。 15

発表が実際に主張していること

老化関連実証の出発点は、バイオメディカル専門家から提供された「aging-relevant genes」のリストである。これをHakkenのエンティティ語彙へ対応させると1,386ノードとなり、959,805の未順序ペアが生じた。THiGERLLMが関係タイプを予測した結果、閾値以上の仮説は1,543,297件となった。つまり「154万個の遺伝子を発見した」のではなく、**約96万ペアに対して複数の関係述語を予測した結果が約154万仮説**である。 ¹⁶

既存グラフにある関係を除くと1,524,046件となり、各エンティティについて上位3候補のみ残すtop-kフィルターで2,804件まで削減された。そこから専門家が信頼度0.8超などを考慮して3仮説を実験へ送った。 ⁹

その3仮説が以下である。

エンティティ	Hakkenが予測した関係	Confidence	Wet-lab判定
SOAT1 → STAT3	affects transcriptional activity of	0.84183	Denied
RAF1 → TNF	decreases expression of	0.83114	Supported
TP53 → BAMBI	affects expression of	0.85391	Supported

原論文Table F6/F7。 ⁹

したがって、ニュースでいう「未知遺伝子関係」は、少なくとも著者らのデータベース上で未記録だった**関係型付きエッジ**を意味する。「未知の遺伝子そのもの」を発見したのではなく、TP53、BAMBI、RAF1、TNFはいずれも既知の遺伝子・タンパク質である。さらに、Hakkenの“novel”判定は基本的に「reference knowledge graphにそのedgeが存在しない」ことを使っているため、「世界中の全論文・全データベースを人間が精査して絶対に一度も示唆されていなかった」という意味より狭い。 ¹⁷

遺伝子の識別

Symbol	標準的な説明	NCBI Gene ID	今回の役割
TP53	tumor protein p53	7157	BAMBI発現に影響すると予測
BAMBI	BMP and activin membrane-bound inhibitor	25805	TP53の予測対象
RAF1	RAF proto-oncogene serine/threonine-protein kinase	5894	TNF発現を低下させると予測
TNF	tumor necrosis factor	7124	RAF1の予測対象

TP53についてはEntrez Gene 7157、HGNC 11998、UniProt P04637、TNFについてはEntrez 7124、HGNC 11892、UniProt P01375という標準識別情報とも整合する。 ¹⁸

報道状況

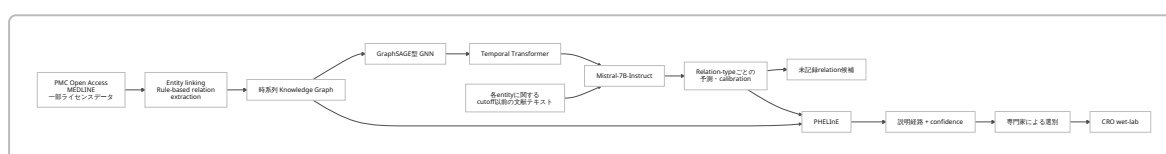
2026年9月8日には日本語AI専門メディアAIDBがHakkenを取り上げ、その後XenoSpectrumも過去文献から未来の発見を予測するシステムとして紹介している。いずれもarXiv発表後数日以内の速報・解説に近く、原論文を独自に追試した報道ではない。 ¹⁹

本調査では2026年9月10日時点で、Nature News、Science、MIT Technology Review等に相当する主要英語科学メディアによるHakken固有の詳細な独立検証記事、またHakkenを名指しして批評する査読済み commentaryを確認できなかった。発表から約1週間しか経っていないため、これは驚くべきことではない。したがって、現時点の報道量やSNS反応を「科学界のコンセンサス」と解釈すべきではない。発表日自体はarXiv記録から確認できる。 20

Hakkenの技術詳細と再現性

全体アーキテクチャ

Hakkenは大きく、**知識グラフ構築 → THiGERLLMによる未知関係予測 → PHEInEによる説明 → 人間専門家による優先順位付け → 実験**という構造を取る。主モデル名THiGERLLMは、前身THiGERにLLMのテキスト意味情報を融合したものと理解できる。 21



上図は原論文のTHiGERLLM/PHEInEおよびwet-lab選択過程を整理した概念図である。 22

データセット

主要な生物医学評価用データは、PMC Open Accessの全文、MEDLINEのタイトル・抄録、および一部のライセンスデータから構成される。文献から最初に約2億4,933万個のraw triplesを得た後、重複、無効エントリ、低頻度関係、矛盾する極性などをクリーニングし、最終的に**254,806エンティティ、7,127,960 triples、23 relation types、20 macro-domains**となった。 23

エンティティの正規化には疾患、遺伝子・タンパク質、化学物質等のオントロジー情報が利用され、relation extractionは完全なend-to-end LLM抽出ではなく、オントロジーへのentity linking、ルールベースNLP、構文簡略化を組み合わせる。各関係は (head entity, relation type, tail entity) のsemantic tripleへ変換される。 24

これは重要な設計判断である。LLMだけに自由文から「関係」を幻覚させる方式に比べ、relation vocabularyを限定し、エンティティをcanonical IDへ落とすため監査しやすい反面、ルールに入っていない表現や条件依存性を落とす可能性がある。実際、論文は**negationとspeculative statementsを現在の抽出データに含めていない**ことを明示的な制約として挙げている。 24

PubTator3自体はNCBI研究者が開発した独立した生物医学文献インフラで、2024年のNucleic Acids Research論文では、約3,600万PubMed abstractsと約600万PMC Open Access full-text articlesを対象に、10億超のentity/relation annotationを提供すると報告されている。したがってオープン版Hakkenの基盤にPubTator3を採用すること自体には相応のデータ基盤上の妥当性がある。 25

THiGERLLM

グラフ側では、各時点の知識グラフに対して近傍情報を集約し、エンティティペアの表現を構成する。原論文のアルゴリズムではGraphSAGE型の近傍集約を行い、それを時間スナップショット列としてTemporal Transformerへ入力する。その結果を複数の「graph tokens」に射影し、LLMのtoken sequenceへ挿入する。 26

LLM側にはMistral-7B-Instruct-v0.3が用いられる。対象エンティティについて、時間カットオフ以前のPMC Open Access文からテキスト文脈を取得し、グラフ情報とテキスト意味情報を融合して23種類のrelation typeをmulti-labelで予測する。モデルはテキストbranchとグラフbranch双方のlogitを生成し、relationごとにensemble係数をvalidation dataで調整する。最終スコアはPlatt scalingとisotonic calibrationを組み合わせで確率へ校正され、relationごとに異なる閾値を使う。²⁷

Positive-Unlabeled Learningが重要な理由

科学文献の知識グラフでは、「関係が書かれていない」ことは「関係が存在しない」と同義ではない。むしろ未発見かもしれない。このためHakkenは0ラベルを通常のnegativeとして強く学習させず、positive-unlabeled (PU) 問題として扱う。論文のDynamically Weighted Balanced Focal lossは、unlabeled negativeの寄与を弱め、既知positiveを重視する。²⁷

これは仮説発見システムとして理にかなう一方、通常のprecisionの意味も曖昧にする。モデルが「false positive」と判定された関係の一部は、単にまだ将来の論文に出現していないだけかもしれないからである。このためHakkenを通常の固定ラベル分類器と同じ感覚で評価するのは危険である。²⁸

バックテスト性能

論文は1990、2000、2010、2020年などを知識カットオフとして、その後に初めて文献へ出現した関係を予測できるかを評価している。1990～2010カットオフを使った長期予測では、予測期間を伸ばすにつれmacro recallは約8ポイント程度低下する一方、ランダムにscoreを割り当てるnull baselineより大幅に高く、nDCGによるランキングは比較的安定した。²⁹

2020年カットオフの代表結果では、THiGERLLMはmacro F1で約51.16、前身THiGERは約50.98と僅差で上回る一方、weighted F1はTHiGER約77.97に対しTHiGERLLM約73.73、nDCGもTHiGER側が高い。THiGERLLMの主な利点は、とくにsupportの少ないrelation typeに対するrecall改善であり、precisionとのtrade-offを伴う。⁵

したがって論文中の「new benchmark」という表現は一定の根拠を持つが、あらゆる評価指標で旧モデルを圧倒したという意味ではない。

PHELIInEによる説明

予測に説明を付与するサブシステムが **PHELIInE (Predicted Hypothesis Elucidation with Literature-Inferred Explanations)** である。候補entity間をつなぐmulti-hop knowledge graph pathを列挙し、GraphSAGE surrogate modelで、ある説明部分グラフだけを残したとき予測scoreをどこまで再現できるか (sufficiency)、説明を除いたときscoreがどれほど落ちるか (necessity) などを評価する。さらに似た経路ばかりを示さないよう、異なる中間ノードを持つ経路をrerankingする。²⁴

ここで著者自身が明確に認める重要な制約がある。**sufficiency／necessityは因果性を示すものではなく、surrogate modelが元モデルをどれだけ忠実に近似するかに依存する**。PHELIInEは「なぜモデルがそう予測したか」をグラフ上で説明する補助であり、「この分子経路が生物学的に因果メカニズムである」と証明するものではない。²⁴

再現性

SonyResearchの公式GitHub **HakkenOSS** は公開されており、knowledge-graph construction、KGE、querying、contextualisation、hypothesis generation、Neo4j、Web UI等を含むpipelineを公開している。source codeはBSD-3-Clauseと表示されている。³⁰

しかし再現性には三段階を区別する必要がある。

再現対象	現状	評価
Hakkenの基本コード・パイプライン	HakkenOSS公開。 ³⁰	比較的高い
PubTator3のみで構築するオープン版	論文でもGUI評価にはライセンス上PubTator3-only版を使用。 ¹⁰	部分的に再現可能
論文主要benchmarkと154万件生成を完全再現	本評価KGはPMC/MEDLINE + select licensed data。 ²⁴	完全再現には障壁
wet-lab raw dataから統計再解析	論文に図、手順、reagent catalog等は詳細掲載。 ³¹	方法は比較的透明だがraw-level data公開は不十分
model parametersの法的状態	GitHub READMEは「CC BY 4.0 or CDLA Permissive 2.0 — pending Sony management confirmation」と表示。 ³⁰	公開直後で未整理部分あり

特に論文の主データと完全なオープン版が同一ではないため、「GitHubが公開された＝論文の全数値を誰でも完全再計算できる」とまでは評価できない。

未知遺伝子関係と実験エビデンス

TP53-BAMBI

Hakkenの予測は **TP53 affects expression of BAMBI**、confidence 0.85391である。関係述語が“increases”または“decreases”ではなく“affects”であるため、予測そのものは方向を限定していない。 ⁹

実験ではHepG2細胞にMDM2阻害剤Nutlin-3を投与してp53 proteinを誘導し、BAMBI mRNAをRT-qPCRで測った。TGF-β1はBAMBIに対するpositive controlとして使われた。最初の24時間実験で10 μM Nutlin-3は、細胞播種密度に応じてBAMBI発現を約17～33%上昇させ、p53 proteinの誘導も確認された。 ³²

変化量が比較的小さかったため、8、16、24、48時間を測定する確認実験が行われた。10 μM Nutlin-3では24時間でBAMBIが約16%、48時間では約38%上昇し、最初の実験と方向的に整合した。 ³²

この結果は「p53経路を薬理的に活性化するとBAMBI transcriptが変化する」という**機能的関連**を支持する。しかし、少なくとも論文で報告された実験には、TP53 CRISPR knockout/knockdown、p53 rescue、BAMBI promoterへのp53結合を示すChIP-seq/CUT&RUN、reporter assay、BAMBI protein定量など、直接転写制御を確定する実験は含まれていない。したがってエビデンスは「TP53→BAMBIの直接転写制御」より一段弱い。 ³³

著者も生物学的解釈については慎重で、将来TP53がBAMBIを抑制すると判明するケースと、逆に活性化すると判明するケースの両方を議論している。BAMBIがTGF-βシグナルの抑制側に働くことから、方向性によってがん、線維化等への意味が異なるためである。つまり原論文自身も、**機構は未確定**と実質的に認めている。

¹⁰

本レポートのエビデンス評価：中程度。

機能的関連の再現性はあるが、直接性・因果方向・老化との関連は未確立である。

RAF1-TNF

Hakkenの予測はより具体的で、**RAF1 decreases expression of TNF**、confidence 0.83114である。 9

THP-1細胞を用い、RAF1選択的阻害剤GW5074とpan-RAF阻害剤AZ628を投与し、TNF mRNAをACTBで正規化してRT-qPCR測定した。RAF1 downstream activityはERK1/2 phosphorylationを代理指標として評価した。LPSはTNF誘導positive controlである。 34

初回実験では、GW5074 1,000 nMでTNF mRNAが4時間でlog2 fold change +1.1、24時間で+1.8と有意に増加した。RAF1がTNFを抑制するならRAF1阻害でTNFが上がることは予測方向と一応整合する。しかし、AZ628では4時間で逆に-1.8~-2.7の低下があり、24時間にはbaselineへ戻った。 34

さらに重要なのはタンパク質レベルである。同じsupernatantで測ったsecreted TNFは、LPSでは658 pg/mLまで上昇した一方、RAF inhibitors群はすべてlower limit of quantification 15.6 pg/mL未満で、**RAF1 inhibitorによるTNF分泌変化は確認されなかった**。 34

このため著者らは確認実験を追加し、別のRAF1 inhibitor ZM336372とGW5074を使用した。GW5074 1,000 nMでは24時間にTNF mRNAが再び+0.9 log2 fold程度上昇したが、secreted TNFはやはり変わらなかった。しかもGW5074およびZM336372処理でpERKが上がり、RAF inhibitorがRAF dimerization/transactivationを通じてMAPK/ERKを逆説的に活性化する**paradoxical ERK activation**と整合する現象が観測された。 32

したがって、ここには非常に重要な意味論的差がある。

支持されたもの：RAF1 pathwayの薬理学的modulationとTNF mRNA regulationの間に何らかの機能的関係がある可能性。

まだ確立していないもの：RAF1 kinase activityそのものがTNF transcriptionを直接かつ単調に抑制する、という因果機構。

阻害剤が狙った通り「RAF1 activityだけを下げた」と単純には解釈できないからである。原論文自身も最終的にはより弱い表現で「potential link between RAF1 signaling and TNF expression」「functional relationship between RAF1 modulation and TNF regulation」とまとめている。 32

本レポートのエビデンス評価：正確な“**RAF1 decreases TNF expression**”については低～中程度、より一般的な“**RAF1/MAPK modulationとTNF regulationの関連**”については中程度。

否定されたSOAT1-STAT3が示すこと

SOAT1-STAT3はconfidence 0.84183で、成功した二候補とほぼ同水準だった。HepG2でLDLによりSOAT1 activityを誘導し、cholesterol ester増加によってtarget engagement自体は確認できたが、STAT3 mRNAにもSTAT3/pSTAT3 proteinにも有意な変化がなく、不支持と判定された。 35

これはHakken評価上、むしろ非常に有用なnegative resultである。confidence 0.84であっても実験が否定されることから、0.8台のscoreを「84%の確率で生物学的に真」と額面通り読むべきではない。Hakkenのconfidenceはモデル内部のcalibrated predictionであり、特定wet-lab protocolでの成功確率そのものではない。 36

エビデンス強度の横断比較

仮説	計算予測	初回wet-lab	確認実験	遺伝学的検証	タンパク質/直接機構	老化表現型	総合評価
TP53-BAMBI	0.85391	BAMBI mRNA +17~33%	24h +16%、48h +38%	未実施	p53 induction確認、直接binding未証明	未実施	中
RAF1-TNF	0.83114	GW5074でTNF mRNA ↑、AZ628では一時 ↓	GW5074の mRNA変化を再現	未実施	secreted TNF変化なし、paradoxical ERKあり	未実施	低~中（正確な方向）／中（広義の関連）
SOAT1-STAT3	0.84183	不支持	—	未実施	pSTAT3/mRNAとも不支持	未実施	否定的

原データの数値と判定は原論文Appendix Fに基づく。 ³⁷

ここからの最重要結論は、「**老化関連**」の部分が現状もっとも弱いことである。各遺伝子が専門家作成の aging-related listに入っていることと、今回新たに予測・検証されたpair relationが実際にaging phenotypeへ因果的に影響することは別問題である。原論文のwet-lab sectionはgene-expression/pathway relationshipを検証しており、寿命やsenescenceそのものを検証していない。 ³⁸

一次データ・追試と専門家評価

一次データの公開状況

原論文Appendix Fは比較的细节で、cell number、投与濃度、時間、biological replicate、RT-qPCR primer/probe、抗体、試薬メーカーおよびcatalog numberまで記載する。RAF1-TNFの主要実験では各条件n=4 biological replicatesとされ、mean ± SEM、vehicle controlとの統計的有意差が表示されている。 ³⁹

一方、本調査で確認した資料では、個々のsampleのraw Ct values、全western/Jess numerical output、plate-level data、完全な解析notebookを別repository accessionとして公開していることまでは確認できなかった。このため図から結果を確認することは可能だが、第三者が原raw dataからすべてのp-valueを独立再計算できる透明性とは区別すべきである。HakkenOSSは主としてソフトウェア／KG／hypothesis-generation infrastructureのリポジトリである。 ⁴⁰

「追試」はあるが「独立再現」はまだない

RAF1-TNF、TP53-BAMBIについては、初回結果が限定的だったため同一研究プロジェクト内でconfirmatory assayを追加している。これは良い研究慣行であり、単発の一条件だけで“confirmed”としたわけではない。

ただしwet-labを担当したのは論文が“independent Contract Research Organization”と呼ぶ外部CROで、謝辞にはPharmagene Discovery Services Ltdとのwet-lab collaborationが記載されている。外部業者による実験という意味ではSony AI内部だけで完結していないが、研究成果は同じHakken論文の検証として設計・報告されている。これは、結果を知らない第三者大学研究室が独自プロトコルで再現して出版する **independent replication**とは異なる。 ⁴¹

2026年9月10日時点では、Hakken自体が9月3日公開の新しいプレプリントであり、TP53-BAMBIまたはRAF1-TNFを第三者がHakken論文後に追試した査読済み論文を確認するには時間的にも極めて早い。 ²⁰

Sony AI内部・協力研究者による評価

研究チームはHakken GUIを7つのacademic/industry organizationのbiomedical researchersへ提供し、qualitative feedbackは「overwhelmingly positive」で、特にhypothesisに説明エビデンスが付随する点が評価されたと報告している。ただし論文から読み取れる範囲では、参加研究者数、blind evaluation、定量的usability score、対照システムとの比較などは示されていない。したがってこれは有望なuser feedbackであるものの、正式なhuman-factors benchmarkではない。 ¹⁰

独立研究者コメントの現状

Hakkenそのものを名指した独立研究者による査読・公開批評は、本調査では確認できなかった。この点は「専門家から反論がないので正しい」ことを意味しない。単に発表から約1週間で評価形成前と考えるべきである。 ²⁰

一方、独立した老化ネットワーク研究を見ると、Gladyshev、Loscalzo、Barabásiらによる2025年の研究は、数千のlongevity-associated genesがhuman interactome上で相互接続するモジュールを形成し、ネットワークベースで老化介入候補を見つける枠組みを提案している。これはHakkenへの直接コメントではないが、**老化は単一遺伝子ではなく関係ネットワークとして扱う必要がある**という基本思想に関してHakkenと整合する独立研究である。 ⁴²

同時に、この独立研究も候補を「experimentally falsifiable」にすることを重視している。したがってHakkenの真の価値はAI scoreの高さそのものではなく、「通常なら人が優先しなかった、検証可能な候補を安価に発見し、実験で落とせる」かどうかにあると見るのが適切である。 ⁴²

本レポートによる専門的評価

論点	肯定的評価	懸念
仮説生成	文献量を人間だけで探索する問題に適している。時間情報も利用。 ⁴³	公表された研究しか学べず出版バイアスを継承
novelty	「現在KGにないedge」を体系的に探索できる。 ⁹	KGにない＝人類が知らなかった、では必ずしもない
検証可能性	relation typeが明示され、wet-labへ落とせる	文脈・濃度・細胞種など条件がrelationに十分表現されない
TP53-BAMBI	2段階実験でmRNA方向を再現。 ³²	薬理学的p53誘導のみで直接制御未証明
RAF1-TNF	GW5074効果を再実験でも観察。 ⁴	別阻害剤との不一致、pERK逆説活性化、secreted TNF陰性

論点	肯定的評価	懸念
aging relevance	aging-related search spaceから候補抽出。 16	aging phenotype自体を測っていない
臨床意義	がん、炎症、薬剤repositioningへの仮説につ ながり得る。10	現段階で治療標的・臨床有効性を主張 するのは早い

最も慎重であるべき表現は「臨床的意義」である。著者らもTP53-BAMBIについて、BAMBI/TGF-βを介してがんや線維化へ影響する**可能性**、RAF1-TNFについて炎症性疾患やがんへの治療的含意の**可能性**を議論しているが、これは実験済みの臨床効果ではなく将来仮説である。10

限界・バイアス・倫理法務

候補選択バイアス

今回もっとも大きな統計的問題は、約154万仮説から3件だけを人間が選んでいる点である。top-kフィルターの後でも2,804件あり、専門家はconfidence > 0.8等を見ながらinterestingnessやbiological relevanceを考慮して3件を選んだ。3

この設計から得られる「2件成功」という結果はproof-of-conceptとして価値があるが、Hakken全体のprecision/positive predictive valueの推定値にはならない。

数学的に、

$$2/3 = 66.7\%$$

ではあるが、それを「Hakkenが生成する仮説の66.7%が正しい」と外挿することはできない。3件はランダム抽出ではないからである。

Relation semanticsが現実の生物学より粗い

Hakkenの現在の表現はconceptとrelationを中心とし、著者自身が将来はquantity、ratio、environmental condition等をKGへ組み込む必要があると述べている。10

RAF1-TNFがまさにその例である。「RAF1 decreases TNF expression」という一文では、細胞種、RAF1 activation state、drug concentration、刺激の有無、4時間か24時間か、mRNAかsecreted proteinかを区別できない。今回の結果が阻害剤、時点、readoutによって異なったことは、**条件なしの単一edgeでは生物学的現象を十分表現できない**ことを示している。4

文献抽出バイアス

Hakkenは科学文献を入力とする以上、「よく研究された遺伝子」「positive result」「人気のある疾患・pathway」がグラフ上で過剰代表される可能性がある。これは論文から導かれる推論であり、個別に定量化されたHakkenの結果ではない。しかし著者自身、現モデルがfactの最初の出現しか表現せず、supporting papers/sentences数、最新観測年、時間的頻度などを将来追加すべきとしている。10

さらにrelation extractionではnegation/speculative statementsを現在含めていない。これにより「AはBを増加させなかった」「AがBを増やす可能性がある」といった否定・不確実性の情報を、肯定edgeと同じ解像度では扱えない。24

時系列リーク

「未来予測」の評価では、2020年より未来の関係を予測する際に2020年以後の情報がモデルへ漏れれば benchmarkが破綻する。HakkenはKGや取得テキストについてcutoffを設けているが、使用したMistral-7Bの pretraining corpusの全時系列を完全に制御できないため、論文はLLM側の残余的なtemporal leakageを完全には排除できない旨を認めている。 ⁴⁴

これはTHiGERLLMとgraph-only THiGERを比較する際にも重要である。LLMが未来文献由来の断片を潜在的に学習済みなら、historic backtestingの一部が純粋な予測能力ではなくpretraining memoryで説明される可能性が残る。現時点では「リークが起きた」と証明されているわけではなく、**完全排除できていないリスク**として扱うべきである。 ⁴⁴

AI説明可能性

PHELInEはブラックボックスモデルに説明経路を与える点では有用だが、著者自身が、surrogate model fidelityへの依存と、causal/counterfactual influenceを直接表現していないことを制約として挙げる。したがってユーザーが説明pathを「AIが解明した分子機構」と読み替えることは危険である。 ²⁴

説明の正しい読み方は、

「この予測scoreをモデル上で支える既存KG経路」

であって、

「自然界で実際にこの経路を通して作用することを実験的に証明した機構」

ではない。

薬理学的検証の限界

RAF1-TNFで特に顕著だが、小分子阻害剤によるperturbationは、target specificity、off-target、feedback、paradoxical activationなどを伴う。実際にRAF inhibitorsでpERK増加が観測され、単純な“RAF1 inhibition”として扱えない結果となった。 ³²

TP53-BAMBIでもNutlin-3はTP53遺伝子を直接編集するのではなくMDM2を阻害してp53を安定化するため、p53に依存する効果とNutlin-induced cellular-state変化を切り分けるには遺伝学的対照が必要である。 ³²

老化モデルとしての外的妥当性

探索対象はaging-related genesであるが、wet-labでは通常ヒト細胞株に短時間～48時間程度の薬理学的処理を行っている。 ⁴

したがって、現時点では以下は未証明である。

TP53-BAMBIまたはRAF1-TNFが加齢に伴い実際に変化すること、老齢個体で若齢個体と異なること、その関係を操作するとcellular senescenceが増減すること、organismal lifespan/healthspanを変えること、ヒトの加齢性疾患リスクを変えること。これらは今回のwet-lab endpointsには含まれない。 ⁴

データ利用・法務

論文主datasetにはPMC OAとMEDLINEのほかlicensed dataが含まれる。これは研究性能を高め得る一方、第三者による完全再配布・再現を制限する。研究者向けGUIをPubTator3-only版にした理由としてもlicensingが挙げられている。 45

GitHubのsource codeはBSD-3-Clauseで公開されている一方、model parametersのlicense欄には調査時点で「CC BY 4.0 or CDLA Permissive 2.0 – pending Sony management confirmation」と記載されている。公開直後特有の未確定事項と見られるため、企業が商用・再配布用途で導入する場合にはlicense text確定を確認すべきである。 30

患者個票を用いた臨床AIではなく公開・ライセンス科学文献と細胞実験が中心であるため、今回の核心的倫理問題は個人プライバシーより、**データprovenance、ライセンス、モデル説明の過剰解釈、AI-generated hypothesis**を「発見」と宣伝する際の責任所在にある。

今後の推奨アクション

最優先は遺伝学的因果検証

TP53-BAMBIについては、薬剤のみではなくTP53 CRISPR knockout/CRISPRi、siRNA、wild-type TP53 rescueを組み合わせ、BAMBI mRNAおよびproteinを測るべきである。p53 wild-type／TP53-nullのisogenic pairでNutlin-3効果が消失するかを確認すれば、p53依存性をかなり明確にできる。

次にp53 ChIP-qPCR、ChIP-seqまたはCUT&RUNでBAMBI promoter/enhancerへの直接占有を確認し、p53-response elementを変異させたreporter assayを行う。この段階で初めて「TP53がBAMBIを直接転写制御する」という機構的主張へ進める。

RAF1-TNFでは、薬理阻害剤のparadoxical activation問題を避けるため、RAF1 CRISPR knockout、inducible degron、CRISPRiを使い、wild-type RAF1とkinase-dead RAF1 rescueを比較する必要がある。BRAF/ARAFを個別に操作する対照を設ければRAF family redundancyも分離できる。今回の薬剤結果が複雑だったことから、これは特に高優先度である。 4

「老化関連」を本当に検証する

現研究が老化関連探索空間からの発見であることを、老化生物学の発見へ格上げするにはモデル自体を変える必要がある。

段階	推奨モデル	測定対象
細胞老化	replicative senescence、DNA-damage-induced senescence、oncogene-induced senescence	p16/p21、SA-β-gal、EdU、SASP、TP53/BAMBI/RAF1/TNF
一次ヒト細胞	若年・高齢donor由来fibroblast、monocyte/macrophage等	pair relationの年齢依存性
組織	若齢・老齢動物の肝臓、免疫系等	expression、protein、pathway activity
因果介入	gene KO/activation + aging model	senescence burden、inflammation、tissue function

段階	推奨モデル	測定対象
個体	マウス等	healthspan、frailty、病態、必要ならlifespan
ヒトデータ	GTEx等の年齢情報を持つ公開omics、aging cohort	age-dependent correlation、eQTL、causal inference

これによって「aging-related genesの間に関係があった」から「この関係がaging biologyを変える」へ主張を進められる。

第三者multi-lab replication

最低2〜3の独立研究室で、事前にprotocol、primary endpoint、effect size threshold、統計手法を登録し、blindに検証するのが望ましい。

TP53-BAMBIなら「TP53操作によるBAMBI mRNA/protein change」、RAF1-TNFなら「genetic RAF1 perturbationによるTNF mRNAおよびsecreted protein change」をprimary endpointにし、探索的条件を後付けして「成功条件だけ選ぶ」ことを防ぐべきである。

真のHakken成功率を測るprospective benchmark

現在の3候補は専門家が選んでいるため、Hakken自身の予測性能と人間のselection skillが混ざっている。これを分離するには、例えば以下のような実験が必要である。

1,000件程度の候補をscore帯ごとに層別化し、Hakken scoreをblindにした実験チームへランダム抽出候補を渡す。0.9以上、0.8-0.9、0.7-0.8などのscore binごとに一定数を検証し、scoreと実験成功率のcalibration curveを作る。同時にhuman expert only、graph-only THiGER、LLM-only、random baselineと比較する。

これにより初めて、

$$P(\text{wet-lab support} \mid \text{Hakken score})$$

を推定できる。

さらに、2026年9月10日時点でモデル・KG・候補リストを凍結し、2027〜2030年に新たに発表される独立論文でどの候補が自然に確認されるかをprospectiveに追跡すれば、「未来予測」という主張に対する最も説得力ある検証になる。

公開すべきデータ

再現性を最大化するには、少なくとも次を公開することが望ましい。

公開物	理由	優先度
1,543,297候補の全relation、score、model version	cherry-picking監査	最高
2,804 top-k候補一覧	wet-lab selection監査	最高
aging gene seed list + canonical IDs	探索空間再現	最高
KG snapshot date / corpus manifest	時系列リーク検証	最高

公開物	理由	優先度
raw qPCR Ct values	統計再解析	最高
raw Jess/Western output	protein evidence再解析	最高
model checkpoints・random seeds	ML再現	高
各hypothesisのPHELIInE paths	explanation監査	高
negative/failed experiments	publication bias抑制	高

論文はaging seed listをSupplementary Informationに置くとしており、HakkenOSSも公開されているため、オープンサイエンス方向への基盤はすでに存在する。 ⁴⁶

推奨共同研究

Sony AI側に最も有益なのは、単なるAI研究機関同士ではなく、**老化生物学、分子機構、薬理学、統計、知識グラフの各専門家を独立に組み合わせたコンソーシアム**である。

老化研究グループにはprimary/aged-cell models、分子生物学者にはCRISPR・ChIP/CUT&RUN、薬理学者にはtarget engagement/off-target解析、統計家にはcandidate samplingとprospective calibration、情報科学者にはKG provenance・temporal leakage監査を担当させる形が望ましい。

特に独立したaging-network研究では、老化関連遺伝子を単一targetではなくinteractome moduleとして捉える方向が進んでいるため、Hakkenの予測edgeを既存のaging interactome、omics、drug-response networkと統合する共同研究は科学的相性がよい。 ⁴²

参考文献とソース優先度

優先ソース一覧

優先度	ソース	種類	信頼度	本調査での用途
S	Besold et al., Hakken: Predicting future discoveries to fill the gaps in today's knowledge, arXiv:2609.04494, 2026-09-03. ⁴⁷	一次論文・プレプリント	非常に高い (研究内容の一次記録)／ 結論確度は査読前	アルゴリズム、データ、benchmark、全wet-lab結果
S	SonyResearch, HakkenOSS. ³⁰	公式コード	非常に高い	OSS範囲、構成、license、再現性
A	Sony AI, Scientific Discovery. ⁴⁸	Sony AI公式	非常に高い	プロジェクト目的・位置づけ
A	Sony AI, “How Sony AI’s Scientific Discovery Team is Reimagining How Researchers Evaluate Hypotheses”, 2026-01-26. ⁴⁹	Sony AI公式ブログ	高い	Hakken以前からの技術思想・研究者中心設計

優先度	ソース	種類	信頼度	本調査での用途
A	Wei et al., PubTator 3.0: an AI-powered Literature Resource for Unlocking Biomedical Knowledge, Nucleic Acids Research, 2024. ²⁵	査読済み 基盤論文	非常に高い	オープン版Hakken の文献データ基盤
A-	Sony AI Publications掲載のAkujuobi et al., Link prediction for hypothesis generation: an active curriculum learning infused temporal graph-based approach. ⁵⁰	関連・前 身研究	高い	THiGER系譜
A-	Besold et al., Literature-based Hypothesis Generation: Predicting the evolution of scientific literature to support scientists, Sony AI publications listing, 2025. ⁵⁰	関連研究	高い	Hakkenの研究系譜
B+	Gross et al., Network-driven discovery of repurposable drugs targeting hallmarks of aging, 2025. ⁴²	独立研究 者による 関連研究	中〜高、プレ プリント	aging network研究 との比較
B	AIDB, Hakken解説、2026-09-08. ⁵¹	日本語専 門メディア	中	日本での初期報道
B-	XenoSpectrum, Hakken解説、2026-09-10. ⁵²	日本語技 術メディア	中	日本語報道・一般向 け解説
C	著者・関係者のSNS/LinkedIn等	一次発信 だが非査 読	低〜中	発表経緯の補助のみ

信頼度の読み方：一次資料であることと科学的結論が確定していることは別である。Hakken arXiv論文は「著者が何を実施・主張したか」を確認するには最高信頼度だが、プレプリントである以上、その主張の妥当性について査読コミュニティの検証はまだ済んでいない。 ²⁰

原論文で特に確認すべき箇所

技術構成は本文およびAppendix B、benchmarkはAppendix C、PHELIInEはAppendix D、Biomedical Data ProvenanceはAppendix E、wet-lab selectionおよび実験詳細はAppendix Fに集中している。とくにTable F6/F7は3仮説の予測と判定、Fig. F8/F9はRAF1-TNF、Fig. F10/F11はTP53-BAMBIのデータを確認するうえで重要である。 ⁵³

原論文の図を見ると、RAF1-TNFについてはFig. F8/F9で阻害剤ごとに異なる挙動が視覚的にも確認できる。これは「Supported」という表中の一語だけを読むより重要な情報である。 ⁵⁴

TP53-BAMBIについてもFig. F10/F11はNutlin濃度・時間依存性を示しており、単純なオン／オフ関係ではないことが分かる。 ⁵⁵

最終判定

本調査時点での科学的ステータスを段階別に表すと、次の評価が妥当である。

命題	判定
HakkenというAI scientific-discovery systemがSony AI主導で開発された	確立。 56
約154万件の老化関連relation hypothesisをscoreした	確立（原論文報告）。 16
TP53-BAMBIがHakkenのKG上で未記録候補だった	確立（原論文定義のnovelty）。 9
p53 pathway activationとBAMBI expressionの機能的関連	細胞実験で支持、追加確認あり。 32
TP53がBAMBIを直接転写制御する	未確定
RAF1 modulationとTNF mRNA regulationの関連	限定的に支持、追加実験あり。 4
RAF1がTNF expressionを直接抑制する	未確定。薬剤間不一致・paradoxical ERK activationあり。 32
両関係が老化を制御する	未証明
老化治療targetとして有効	仮説段階
Hakkenの仮説成功率が約2/3	支持できない。選択バイアスが大きい。 9
第三者研究室が独立追試済み	2026年9月10日時点で確認できず
peer reviewを通過済み	いいえ。現時点ではarXivプレプリント。 20

総合すると、Hakkenの今回の成果は「老化そのものの新しい分子機構を確定したブレークスルー」ではまだないが、「文献から未来のrelationを予測するAIが、少なくとも人間が実験可能と判断できる新規候補を生成し、そのうち二つで再現可能な分子応答を観測した」というproof-of-conceptとしては注目に値する。特にTP53-BAMBIは遺伝学的・転写制御実験へ進める価値が高く、RAF1-TNFはむしろ今回の複雑な薬理結果を解きほぐすこと自体が重要な次段階になる。 57

Hakkenの真価を決める次のマイルストーンは、派手な候補数ではなく、公開した全候補を固定したprospective test、ランダム化されたcandidate validation、独立multi-lab replication、遺伝学的因果実験、そして実際のaging phenotypeでの再現である。そこまで到達すれば、「文献検索を便利にするAI」から「新しい科学知識を前向きに予測するAI」へ一段強い証拠を得ることになる。

1 7 13 20 47 56 Hakken: Predicting future discoveries to fill the gaps in today's knowledge
https://arxiv.org/abs/2609.04494?utm_source=chatgpt.com

2 3 4 6 8 9 10 11 16 17 21 22 23 24 26 27 29 31 32 33 34 35 36 37 38 39 41 44 45
46 53 54 55 57 Hakken: Predicting future discoveries to fill the gaps in today's knowledge
<https://arxiv.org/pdf/2609.04494>

5 28 Hakken: Predicting future discoveries to fill the gaps in today's knowledge
<https://arxiv.org/abs/2609.04494>

12 14 48 AI-Augmented Scientific Discovery | Sony AI Research

<https://ai.sony/research-areas/scientific-discovery?hsLang=en>

15 50 Sony AI Publications | Research Papers & Scientific Findings

<https://ai.sony/publications?hsLang=en>

18 P53

https://fr.wikipedia.org/wiki/P53?utm_source=chatgpt.com

19 51 科学的発見を予測するAIシステムHakken - AIDB

https://ai-data-base.com/archives/tanshin/tanshin-2097133897498951822?utm_source=chatgpt.com

25 PubTator 3.0: an AI-powered Literature Resource for Unlocking Biomedical Knowledge

https://arxiv.org/abs/2401.11048?utm_source=chatgpt.com

30 40 GitHub - SonyResearch/HakkenOSS: Open Source version of the project Hakken · GitHub

<https://github.com/SonyResearch/HakkenOSS>

42 Network-driven discovery of repurposable drugs targeting hallmarks of aging

https://arxiv.org/abs/2509.03330?utm_source=chatgpt.com

43 49 How Sony AI's Scientific Discovery Team is Reimagining How Researchers Evaluate ... - Sony AI

<https://ai.sony/blog/how-sony-ais-scientific-discovery-team-is-reimagining-how-researchers-evaluate-hypotheses?hsLang=en>

52 過去の文献から未来の発見を導く、Sony AIの仮説生成 ...

https://xenospectrum.com/sony-ai-hakken-knowledge-prediction-preprint/?utm_source=chatgpt.com