

厚生労働行政推進調査事業費（化学物質リスク研究事業）
トキシコゲノミクスとシステムバイオロジーとの融合による
新型化学物質有害性評価系の実装研究
（21KD2001）

令和5年度 分担研究報告書

システムバイオロジーによる毒性解析のAI化

研究分担者 北野 宏明

特定非営利活動法人 システム・バイオロジー研究機構
会長

研究要旨

システム毒性学的手法として、大規模データベースから多くの情報を効率良く抽出し、それによってより規模の大きな網羅的解析を実施すべく、深層学習(Deep Learning)を用いて、膨大なデータから、生物学的に有意な変化を呈した遺伝子を高精度且つ自動的に抽出する自動深層機械学習システムの開発を進めた。特に全ゲノムに及ぶヒストン修飾及びDNAメチル化の解析画像を自動判別する技術開発を進めた。また転写制御解析ツール SHOE の改良を実施し、遺伝子発現データとヒストン修飾などエピゲノム解析における連携強化を進めた。

研究協力者

長谷 武志 特定非営利活動法人

システム・バイオロジー研究機構

Natalia Polouliakh 株式会社ソニーコンピュータ
サイエンス研究所

実施し、ツール間連動を強化することで、高度なAI駆動型システム毒性学基盤の構築を推進する。

A. 研究目的

システム・レベルで毒性を理解するには、膨大な実験データを格納したデータベース、文献、数値モデルなどを統合的に解析する必要があり、大規模かつ複雑なデータを意味のある形で解析するには、深層学習やテキストマイニングなどを含めた一連の人工知能(AI)アルゴリズム群の連携が有効である。さらに、複数の解析ツールをスムーズに連動させる必要がある。本分担研究では、一連の解析過程のAI化を

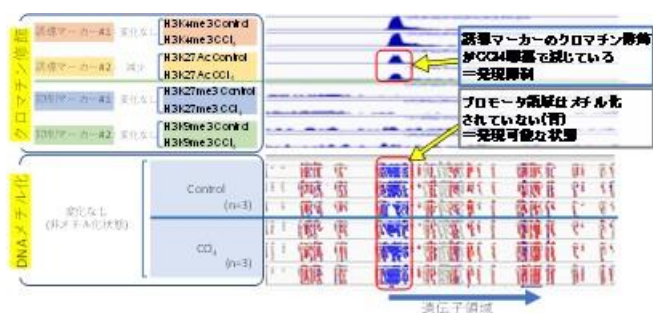
B. 研究方法

システム・レベルで毒性を理解するには、膨大な実験データを格納したデータベース、文献、数値モデルなどを統合的に解析する必要があり、大規模かつ複雑なデータを意味のある形で解析するには、深層学習やテキストマイニングなどを含めた一連の人工知能(AI)アルゴリズム群の連携が有効である。本分担研究では、一連の解析過程のAI化を実施する。

●深層学習を用いたエピゲノム解析データからの有意なエピゲノム修飾の判別

複雑な遺伝子制御ネットワークのかく乱により引

き起こされる化合物毒性のメカニズムの解明には、反復投与（または単回暴露）により引き起こされる“クロマチン修飾とメチル化（エピゲノム修飾）”に関する情報が必要不可欠である。本研究では、国立医薬品食品衛生研究所・安全性生物試験研究センター・毒性部から提供を受けたエピゲノム修飾に関するデータ（エピゲノムデータ）を用いて、深層学習を基盤とした解析システムの構築を進めている。



図A. エピゲノム修飾に関する画像の例
画像の上部に描画されている分布がクロマチン修飾の状況を表しており、下部に描画されているプロットがメチル化の状況を表している。横軸はゲノム配列における位置を表している。メチル化については、青いプロットがメチル化されていない状態を表している。

このデータでは、化合物の反復投与（および単回暴露）により引き起こされる遺伝子のエピゲノム修飾を、図Aのような画像として表現している。このような画像をこれまで、長年経験とトレーニングを積んできた研究者が目で見ても検出し、エピゲノム修飾が遺伝子発現に影響を与えているかどうかの分類を行ってきた。しかしながら、個々の化合物の反復投与と実験に対して遺伝子数と等量の画像が存在するため、解析には多大な時間と労力が必要な状況であった。

このような問題を解決するために、本研究課題（令和3年度から5年度）において、我々は、様々な画像解析のタスクにおいて高い性能が報告されている“深層学習”を基盤とした、エピゲノムデータの解析システムの構築を進めている。深層学習を用いることにより、専門家の目視での情報の統合および判断を行っている工程について、偏奇のない網羅的且つ高速な自動化を行うことが出来る可能性がある。

提供を受けたエピゲノム画像データは、四塩化炭素、バルプロ酸ナトリウム、クロフィブラートの新型反復暴露と単回暴露において得られたものである。各遺伝子に対するエピゲノム画像を、専門家が視覚的に検証して、以下の3群に分類している。

- ① **suppression 群**：反復暴露によるエピゲノム修飾により、遺伝子発現が抑制されたもの（5,937 画像）
- ② **induction 群**：反復暴露によるエピゲノム修飾により、遺伝子発現が誘導されたもの（457 画像）
- ③ **non significant 群**：反復暴露によるエピゲノム修飾により、遺伝子発現が有意な変動をしめさなかったもの（2,349 画像）

この3群の内、化合物の毒性メカニズムを探索するために有意な意味の有る群は **suppression 群** と **induction 群** である。

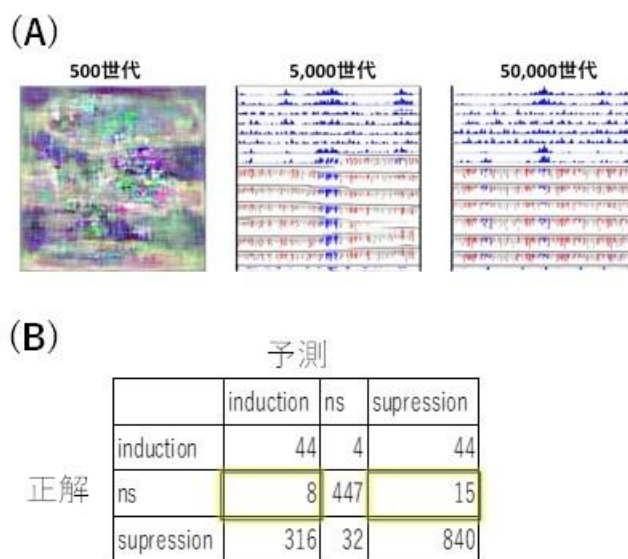
本研究課題において、初年度（令和3年度）は、提供を受けたエピゲノム画像データを訓練データとして用いて、8種の畳み込み層を基盤とした深層学習アーキテクチャ（CNN, convolutional neural network）に基づく深層学習モデルを構築し、上記の3群（**induction 群**、**suppression 群**、および、**non-significant 群**）の予測精度の評価を行った。具体的には、エピゲノム画像データの80%をトレーニングデータとして用いて深層学習モデルを構築し、残りの20%のデータをテストデータとして用いて構築したモデルの分類精度の検証を行った。分類精度の検証の結果、構築した8種の分類モデル全てで、**non-significant 群**と、それ以外の2群を正確に分類できることを確認できた（図B）。この結果は、構築した深層学習モデルは、反復暴露により生じるエピゲノム修飾の中でも遺伝子変動に影響を与えるものを見分けることが出来る可能性を示唆している。しかしながら、構築した8種類のモデル全てで、**suppression 群**と、**induction 群**を分類することが出来なかった。これは、**induction**

群に対する訓練用画像の枚数が少なく、情報が十分に学習できていないことが原因であると考えられた。また、大量の訓練画像の生成のためには、多大な専門家の労力と時間が必要であるため、induction 群の訓練画像の増幅が難しい状況であった。

		予測		
正解		→		
		induction	ns	supression
Resnet18				
	induction	0	3	89
	ns	0	466	4
	supression	0	4	1184
Resnet34				
	induction	0	2	90
	ns	0	468	2
	supression	0	3	1185
Resnet50				
	induction	0	2	90
	ns	0	465	5
	supression	0	3	1185
Alexnet				
	induction	0	2	90
	ns	0	461	9
	supression	0	4	1184
Densnet121				
	induction	0	3	89
	ns	0	466	4
	supression	0	2	1185
Sqeeznet1.0				
	induction	0	1	91
	ns	0	465	5
	supression	0	4	1184
vgg16				
	induction	0	3	89
	ns	0	467	3
	supression	0	3	1185
vgg19				
	induction	0	1	91
	ns	3	464	3
	supression	0	3	1185

図B. 代表的な CNN による分類性能の検証結果

そこで、本研究課題の2年目（令和4年度）は、以上の問題に対処するために、画像の生成モデル（Generative adversarial network : 1. Bingchen Liu, Yizhe Zhu, Kunpeng Song, Ahmed Elgammal (2021) Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. arXiv:2101.04775）の実装を行い、訓練画像(induction 群)の増幅を行った。生成モデルにより、実際の画像と見分けが付かない程度の、高品質な画像の生成に成功した(図C-A)。さらに、CNNに加えて、vision transformer を基盤とした深層学習アーキテクチャの実装も行った。増幅した訓練画像を使用して、代表的な深層学習アーキテクチャを基盤とした分類モデルの構築を行い、その予測精度について検証を行った。しかしながら、増幅した訓練データを用いて構築した深層学習モデルであっても、induction 群、と、supression 群を見分けることは困難であった。特に、ほぼ完全に分類できていた non-significant 群についても、supression 群、または、induction 群と誤判定するものが生じてしまった(図C-B)。



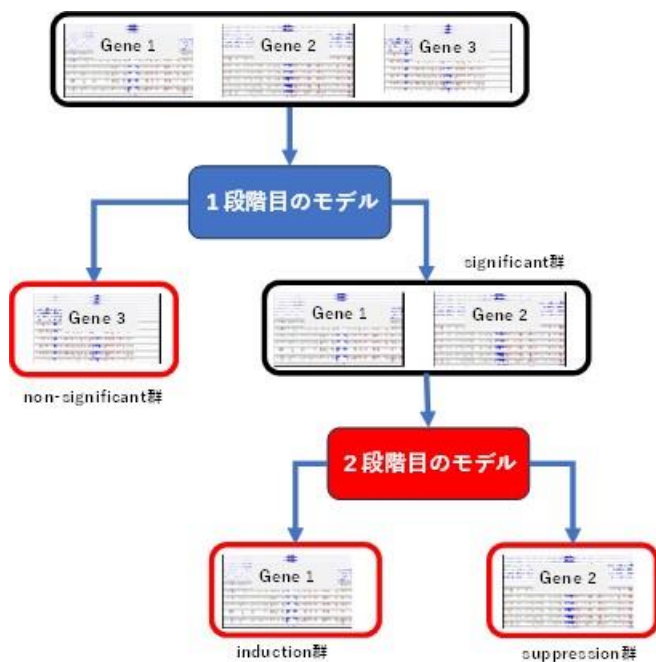
図C. 生成A Iによる画像の増幅と増幅画像を訓練データに追加した場合の予測精度の検証結果。(A) 生成A Iにより生成された画像 (B) 予測精度 (vision transformer)

上記の non-significant 群の誤判定の問題は、令和3年度に作成した増幅したデータを用いないモデルで non-significant 群を正確に分けることが出来たことから、以下の様に2段階のモデルを構築することで解決できる。

1段階目のモデル：non-significant 群とそれ以外の群を分類する深層学習モデル

2段階目のモデル：induction 群と supression 群を二値分類する深層学習モデル

具体的には、図Dに示すように、まず、1段階目のモデルで non-significant 群を分類し、残りの画像を、2段階目のモデルで、induction 群、supression 群に分類する。この様にすることにより、non-significant 群を、induction 群または supression 群に誤判定することを避けることが出来る。



図D. 2段階モデルによるエピゲノムデータの分類フレームワーク

そこで、今年度は、この2段階モデルの構築を行った。また、Induction 群、suppression 群を分類する2段階目のモデルについては、より高精度の分類を達成することを目的として、最新の CNN を基盤としたアーキテクチャ (ConvNeXt)、および、transformer を基盤とした最新のアーキテクチャ (SwinTransformer、MaxVit) の実装も行った。さらに、構築した深層学習モデルが、画像のどの部分に着目して分類を行っているかを調べるために、最新の explainability model (XGrad-Cam : 2. Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, Biao Li (2020) Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. arXiv:2008.02312)を用いて深層学習モデルの解析を行った。

●転写領域解析ソフトウェア SHOE の改良

SHOE の開発は、Java 言語 (USA, Oracle Inc.) で行った。Garuda Platform 用ソフトウェア (Garuda ガジェット) の開発や他の Garuda ガジェットとの連動については、GarudaDevPack を使用した。

性能評価や試験運用には、Perccellome データベースより実際の化学物質曝露による遺伝子発現時系列データを用いた。

C. 研究結果

●深層学習を用いたエピゲノム解析データからの有意なエピゲノム修飾の判別

エピゲノム解析画像を①suppression 群、②induction 群、③non significant 群の三群に分類することを目的として、代表的な CNN および transformer を基盤とした、2段階モデル (図D) によるエピゲノムデータの解析フレームワークを構築した。

図Dに示すように、1段階目のモデルでは、non significant 群とそれ以外の群 (significant 群) を分ける2値分類を行う。そこで、1段階目のモデルについては、induction 群と suppression 群をまとめて significant 群とし、Resnet-18 アーキテクチャを用いて構築を行った。1段階目のモデルについては、エピゲノム画像データの内、モデルの構築に使用をしていない残りの 20%のデータをテストデータとして用いて分類精度の検証を行った。図Eに示すように、1段階目のモデルは、テストデータに対して、正確に non significant 群、significant 群を分類できることが示された。

		予測	
		non significant	significant
正解	non significant	470	0
	significant	3	1277

図E. 1段階目のモデルの予測精度の検証結果

2段階目のモデルでは、1段階目で significant 群と判定されたデータを、induction 群、または、suppression 群へと、2値分類する。2段階目のモデルを構築するために、図Fに示す深層学習アーキテクチャを実装した。実装には、python の代表的な深層学習用のモ

ジュールである、pytorch、timm、および、torchvisionを用いた。

アーキテクチャ	基盤技術	発表年度	被引用回数	参考文献
Resnet-18	CNN	2015	195932	https://arxiv.org/abs/1512.03385
ConvNext	CNN	2022	3056	https://arxiv.org/abs/2201.03545
Vision transformer	transformer	2020	27687	https://arxiv.org/abs/2010.11929
Swin transformer	transformer	2022	949	https://ieeexplore.ieee.org/document/9879380
MaxVt	transformer	2022	268	https://arxiv.org/abs/2204.01697

図F. モデル構築に使用した深層学習アーキテクチャ 被引用回数については google scholar の情報に基づき記載した

2段階目のモデルについては、エピゲノム画像データの80%、および、生成AIにより生成した画像データを用いて構築を行い、残りの20%のデータをテストデータとして用いて構築したモデルの分類精度の検証を行った。図Gに示すように特に、5つのモデルの中で、Resnet-18が最も高い分類性能を示したが、induction群と suppression群を十分に正しく分類することはできなかった。この結果は、エピゲノム修飾データの全体画像を用いた場合、significant群の中で suppression と induction を正確に分類することは、困難であることを示唆する。

Resnet-18			ConvNext		
	induction	suppression		induction	suppression
induction	31	61	induction	13	79
suppression	128	1060	suppression	148	1040

Vision transformer			Swin transformer		
	induction	suppression		induction	suppression
induction	33	59	induction	92	0
suppression	296	892	suppression	1188	0

Max transformer		
	induction	suppression
induction	0	92
suppression	0	1188

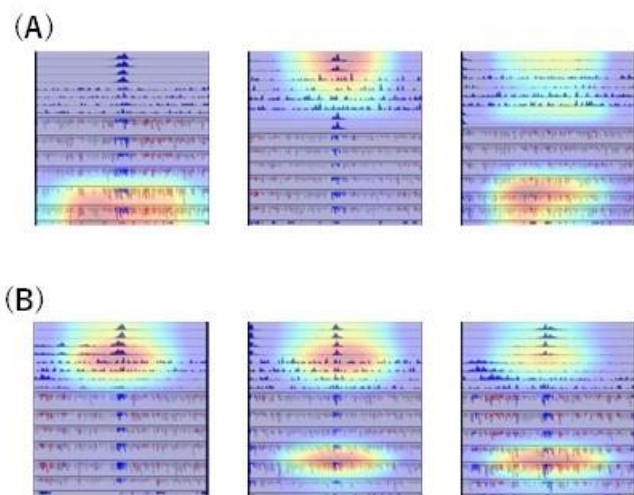
予測

正解

図G. 2段階目のモデルの予測精度の検証結果

さらに、1段階目のモデルについて、

Non_significant群と significant群を分類する判断根拠を探索するため、最新の explainability model (XGrad-Cam) を実装し解析を行った。図Hに示すように、深層学習モデルは、図中の赤および黄色の部分に着目して分類を行っている。Significant群を分類する際には、クロマチン修飾のみに着目している場合(図H-Aの中央図)、メチレーションのみに着目している場合(図H-Aの左図)、そして、クロマチン修飾とメチレーションの両者に着目している場合が存在した(図H-Aの右図)。一方、non-significant群を分類する際には、クロマチン修飾のみに着目している場合(図H-Bの左図)、そして、クロマチン修飾とメチレーションの両者に着目している場合が存在した(図H-Bの中央図、右図)。以上の結果は、エピゲノム修飾のうち、クロマチン修飾のみが遺伝子発現に影響が及ぼされる場合、および、メチレーションのみが遺伝子発現に影響を与える場合、が存在する可能性があることを示唆する。

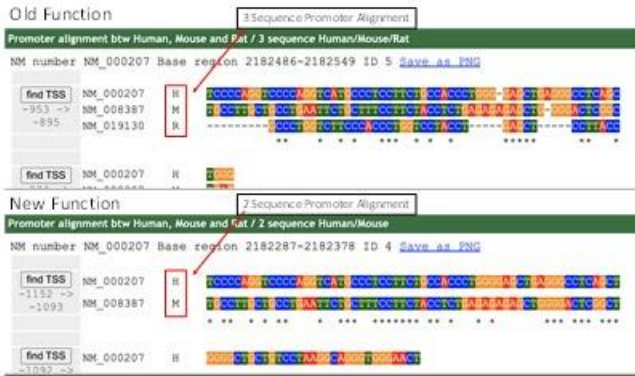


図H. 1段階目の深層学習モデル (significant群と non-significant群を分類するモデル) の判断根拠の可視化 (A)significant群に対する判断根拠の可視化結果 (B)non-significant群に対する判断根拠の可視化結果

●転写領域解析ソフトウェア SHOE の改良

今年度は、SHOEの機能強化を継続し、その性能を評価した。具体的には、当初 Human/Rat/Mouse の

全てでプロモーター配列が保存されていないと候補から落とされていたところを、Human/Mouse 若しくは Human/Rat で保存されていれば候補に残すようにした。



SHOE Updated Feature: 2-Sequence Alignment



Human/Mouse 若しくは Human/Rat の 2 種アライメントに対応した新機能画面

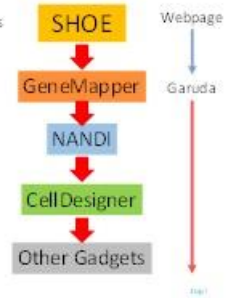
また Garuda Platform 上で、Perccellome DB の利用に適した解析パイプラインの構築を進め、SHOE と他の Garuda Gadget との連携も強化した。



Garudification

GeneMapper is an UI that pulls data from SHOE and sends it to Garuda

- Current Garudification design is simpler and less complex to implement
- GeneMapper connects to SHOE to pull data
- SHOE is independent to Garuda
- GeneMapper is connected to Garuda and can send SHOE's data to other Gadgets **Discover**

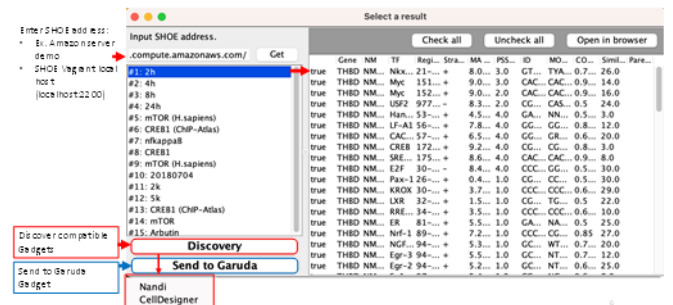


Garuda 対応と解析パイプラインの流れ

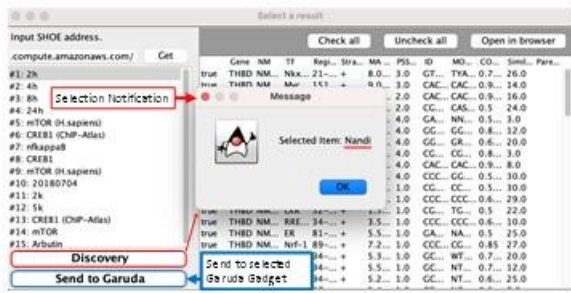
実際のパイプラインの流れは以下の通り。



GeneMapper UI



GeneMapper UI

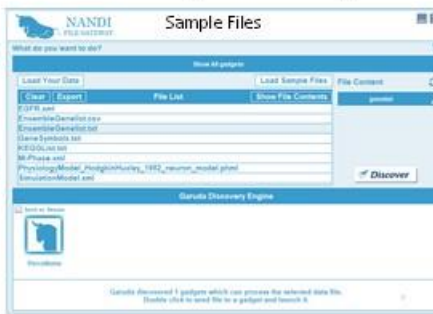


Send to NANDI | SHOE data → GeneMapper → NANDI → Gadget

NANDI loads data and sends it to other compatible Gadgets

GeneMapper and NANDI are connecting

- Issue to correct: Data not being attached when sent

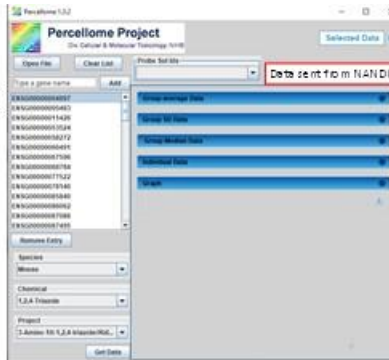


Percellome

SHOE data → GeneMapper → NANDI → Percellome (others)

Example:

- Connecting to Percellome Gadget



実際のデータで解析を試みて、Garuda Gadget 運動による解析パイプラインの性能や使い勝手の評価を進めた。

D. 考察

本研究課題で行ったエピゲノム解析において、深層学習モデルは、反復曝露によるエピゲノム修飾の内、遺伝子発現に影響がある修飾を分類できる可能性が示唆された。しかしながら、エピゲノム修飾の全体画像を用いた場合、induction 群、suppression 群

の分類は困難であることが分かった。また、最新の explainability model を使用した深層学習モデルの解析を行ったところ、深層学習モデルでは、遺伝子発現に影響がある修飾を分類するときに、エピゲノム修飾画像の中の、(i)クロマチン修飾に関する画像部分のみに着目して分類している場合、(ii)メチレーションに関する画像部分のみに着目して分類を行っている場合、があることがわかった。これらの結果は、エピゲノム修飾のうち、クロマチン修飾のみが遺伝子発現に影響が与える場合、および、メチレーションのみが遺伝子発現に影響を与える場合、が存在する可能性があることを示唆する。そこで、suppression 群と induction 群をより正確に分類することを目的として、今後、①画像のクロマチン修飾部分のみ (図Aの画像の上半分) を用いて、suppression と induction の分類を行うモデルの構築、および、②画像のメチレーション部分のみ (図Aの画像の下半分) を用いて、深層学習モデルの構築を進める予定である。

一方、転写領域解析パイプラインの改良については、SHOE の機能追加により解析対象の候補数が増加した。また Garuda 化も終え、他のガジェット(解析ソフトウェア)との連携も蜜となり、より実用的となった。

E. 結論

先行研究で開発した遺伝子発現解析用 AI の論文を発表した。

エピゲノム解析の AI 自動化の試みでは、AI が注目している要素を explainability model を利用することで明らかにするなど、深層学習が大規模データの自動分類に有効であることが示された。また AI が注視する領域が判明したので、今後は重要な領域別のモデル構築を進める。

転写領域解析ソフトウェア SHOE については、機能強化の結果、実際の遺伝子発現解析でも利用しやすくなった。SHOE の Garuda 対応も終え、他の

Garuda Gadget との連携もシームレスに実行できるようになった。

F. 研究発表

1. 論文発表

- (1) Takeshi Hase, Samik Ghosh, Ken-ichi Aisaki, Satoshi Kitajima, Jun Kanno, Hiroaki Kitano, Ayako Yachie. DTox: A Deep neural network-based in visio lens for large scale Toxicogenomics data. J. Toxicol. Sci. 2024; 49(3): 105-115.[doi.org/10.2131/jts.49.105].
- (2) González-Pech RA, Li VY, Garcia V, Boville E, Mammone M, Kitano H, Ritchie KB, Medina M. The Evolution, Assembly, and Dynamics of Marine Holobionts. Ann Rev Mar Sci. 2023 Aug 8. doi: 10.1146/annurev-marine-022123-104345. Epub ahead of print. PMID: 37552896.
- (3) Mori T, Takase T, Lan KC, Yamane J, Alev C, Kimura A, Osafune K, Yamashita JK, Akutsu T, Kitano H, Fujibuchi W. eSPRESSO: topological clustering of single-cell transcriptomics data to reveal informative genes for spatio-temporal architectures of cells. BMC Bioinformatics. 2023 Jun 15;24(1):252. doi: 10.1186/s12859-023-05355-4. PMID: 37322439; PMCID: PMC10268514.

2. 学会発表

- (1) 北野宏明「AIがもたらす科学の発見」LINK-J & デンマーク王国大使館主催 ライフサイエンスセミナー『データがつくる生命科学』～AIと時系列データが生み出すパワー、(2023.6.27)、東京、招待講演
- (2) 長谷武志「深層学習を基盤としたオミックスデータ解析とその創薬・毒性学研究への応用」2023年日本バイオインフォマティクス学会年会・第12回生命医薬情報学連合大会、(2023.9.9)、千葉、招待講演
- (3) 長谷武志「AI駆動型の創薬・生命・医学研究の最新動向」AI創薬・ケモインフォマティクスセミナー、(2023.11.24)、オンライン、招待講演
- (4) 北野宏明「Nobel Turning Challenge～Creating the Engine for Scientific Discovery～」キヤノン財団リユニオン2023、(2023.12.15)、東京、特別講演
- (5) 北野宏明「複雑なシステムを理解するサイエンスとテクノロジー：AI駆動型システム生物学への展開」第59回日本東洋心身医学研究会、(2024.3.9)、東京、特別講演

G. 知的所有権の取得状況

1. 特許取得

なし

2. 実用新案登録

なし

3. その他

なし