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 4 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

システム毒性では、一連の解析手順の高度な連携と同時に、大規模データベースから多くの情報を抽出し、それを解析へと結びつける必要がある。本研究では、深層学習(Deep Learning)を用いて膨大な遺伝子変動データから有意に変動した遺伝子を高精度で自動同定させる技術ならびに解析パイプラインの連動強化を行なった。

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

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。

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。

●深層学習を用いた大規模遺伝子発現データベースからの重要遺伝子群の判別

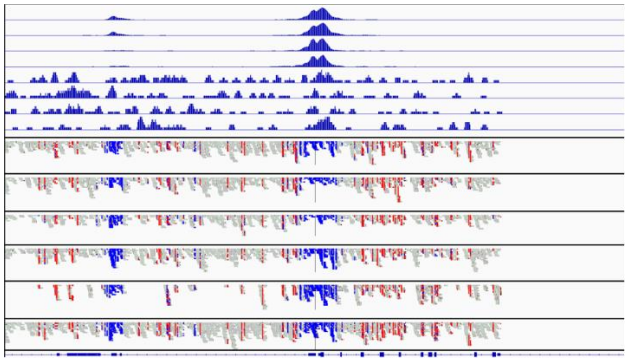
先行研究で開発した、深層学習を用いた３次元グラフの画像解析システム DTOX について改良、即ち特異パターンの追加学習と GUI 実装の改良を進めた。

追加学習用の画像セットは、遺伝子発現を用量×時間×発現量（Percellome 法により細胞 1 個あたりの mRNA コピー数に換算したデータ）の 3 次元グラフに描画したものを用いた。また GUI 実装は、python の代表的な GUI 作成用のモジュールである、PYQT5

と、Qtdesigner を用いて改良を進めた。

●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

エピゲノム解析では、曝露下の遺伝子のエピゲノム修飾（ヒストン修飾およびゲノム DNA メチル化の状態）を、下図に示すような解析画像として表示し、研究者の判断を助けている。



図．エピゲノム解析画像の例
上から８行で示される分布がクロマチン修飾の状況を表しており、下から８行のプロットがメチル化の状況を表している。横軸は配列の位置を表している。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。

有意なエピゲノム修飾を同定するために、長年の経験を積んできた研究者が、それぞれの遺伝子に対するエピゲノム解析画像を検討し、分類を行ってきた。しかしながら、化学物質数×遺伝子数の解析画像が存在し、網羅的に有意なエピゲノム修飾を同定するには、多大な時間と労力が必要となっている。これを解決すべく、エピゲノムデータ（解析画像）から、効率良く有意なエピゲノム修飾を判別する手法の構築を目的として、深層学習モデルの構築を行った。

深層学習モデルの訓練には、専門家により分類されたエピゲノム解析画像を訓練データとして用いた。この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曝露と単回曝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解析画像を、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：反復曝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：反復曝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：反復曝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 群であるが、induction 群の画像の枚数が少なく、最も分類が難しい状況であった。

今回、この induction 群に対する訓練画像を増幅することを目的として、深層学習を基盤とした代表手的な生成モデルである“Generative adversarial network (GAN)”の実装を行った(参考文献 1)。

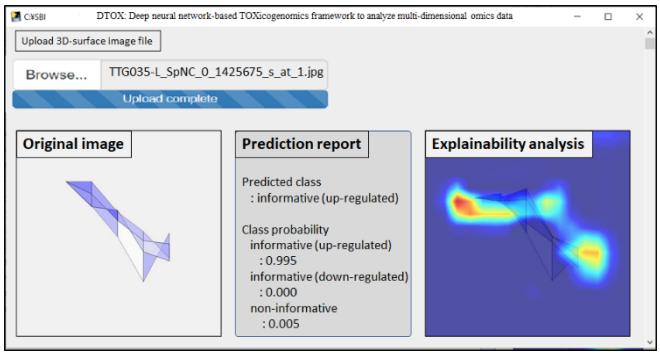
●転写領域解析ソフトウェア SHOE の改良

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、Percellome 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

●深層学習を用いた大規模遺伝子発現データベースからの重要遺伝子群の判別

先行研究で開発した深層学習モデル DTOX については、グラフィカルユーザーインターフェイス（GUI）を実装し、試用とフィードバックによる改良を継続している。

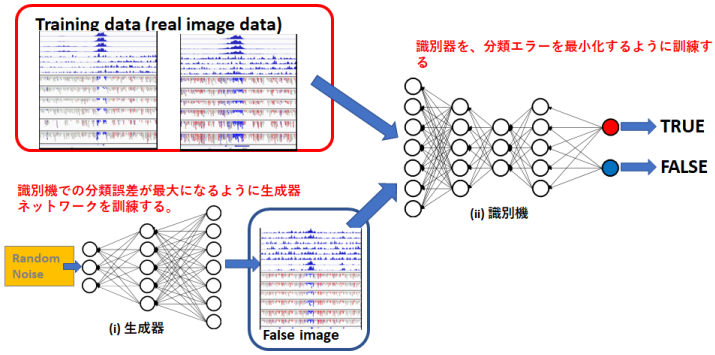


また DTOX に関する成果の原著論文化を進めている。

●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

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、先行研究では多様な深層学習アーキテクチャに基づく分類モデルを構築した結果、構築した分類モデル全てで、non-significant 群と、それ以外の 2 群を正確に分類することが出来たが、induction 群の画像が少なく、分類が難しい状況であった。

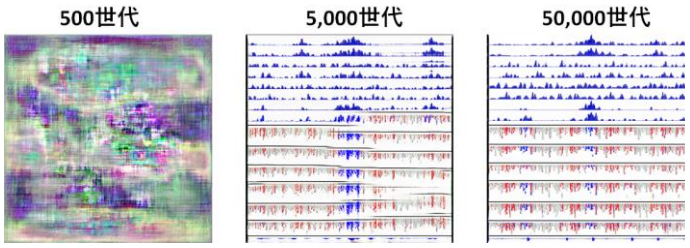
今年度はこの induction 群に対する訓練画像を増幅することを目的として、深層学習を基盤とした代表的な生成モデルである“Generative adversarial network (GAN)”の実装を行った。GAN は、下図に示すように、(i)生成器ネットワークと(ii)識別器ネットワークの 2 つのネットワークから構成されている。生成器ネットワークは、ランダムなノイズから偽の画像データを生成する。識別器は偽の画像データと実物の画像データを比較する。比較により、生成器ネットワークによる画像データから、偽の画像を識別する。生成器ネットワークと識別器ネットワークが競い合う様に訓練が進んでいき、最終的に、GAN は実際の画像データと同様の特徴を持つ新規の画像を生成可能となる。しかしながら、GAN 自体を訓練するためにも、多くの画像データが必要であるという問題があった。そこで、今回は、GAN の改良版である“Fast GAN”の実装を行った(参考文献 3)。Fast GAN では、識別器ネットワークに AutoEncoder を組み込んだ構造を採用している。その構造の中で、中間層の出力から元の画像を復元する Decoder を追加して、その復元誤差をロスに加えて効率の良い学習を実現しており、少ない訓練画像から、訓練画像と同様な性質を持つ画像を生成出来る。



GAN の構造

今回、昨年度と同様に、我々はエピゲノム画像データを分類することを目的として、多様な深層学習アーキテクチャに基づく分類モデルを構築した。これらの深層学習アーキテクチャは、代表的な大規模画像データセット（1000 カテゴリーに分類できる 120 万枚の画像で構成される）である ImageNet dataset を用いてプレトレイニングされたものである。これらを、エピゲノム画像に対して転移学習を行い、分類モデルを構築した。induction 群に対しては、訓練データとして、専門家から提供を受けた 457 枚の実データ画像に加えて、今年度実装を行った生成モデルで生成した画像 6,000 枚も、訓練データとして用いた。

実装した Fast GAN を使用し、induction 群の画像を生成するモデルの構築を行った。今回、生成モデルについて、500 世代、5,000 世代、50,000 世代の訓練を行い、その生成画像の特性について検証を行った（下図）。



Fast GAN による生成画像

図に示すように、500 世代では、ほぼ訓練画像の特徴を捉えられていない。5,000 世代では、訓練画像の主な特徴を捉えられてはいるものの、不鮮明な部分が存在しているが、50,000 世代では、ほぼ完全に訓練画像の特徴をとらえた画像を生成することができた。この結果は、この生成モデルを用いることにより、訓練画像データの生成のために必要であった専門家の労力と時間を、大幅に削減できる可能性があることを示している。

次に、50,000 世代のモデルを使用して、6,000 枚の induction 群の生成画像を生成し、深層学習モデルの訓練に用いた。

今回、エピゲノム修飾に関する実データ画像の内の 80%と、生成モデルで生成した画像を、トレイニングデータとして用いて深層学習モデルを構築した。その後、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（下図)。図に示すように、構築した 7 種の分類モデル全てで、non-significant 群と、それ以外の 2 群を正確に分類することが出来た。しかしながら、構築した 7 種類のモデル全てで、induction 群を分類することが困難であった。

Resnet18

	induction	ns	supression
induction	0	1	91
ns	0	469	1
supression	0	3	1185

Alexnet

	induction	ns	supression
induction	0	2	90
ns	2	462	6
supression	2	13	1173

Resnet34

	induction	ns	supression
induction	0	2	90
ns	0	468	2
supression	0	2	1186

Densenet

	induction	ns	supression
induction	1	1	90
ns	0	467	3
supression	0	2	1186

Resnet50

	induction	ns	supression
induction	0	1	91
ns	0	465	5
supression	0	2	1186

Squeeznet

	induction	ns	supression
induction	0	2	90
ns	0	467	3
supression	0	4	1184

正解
予測

Vgg18

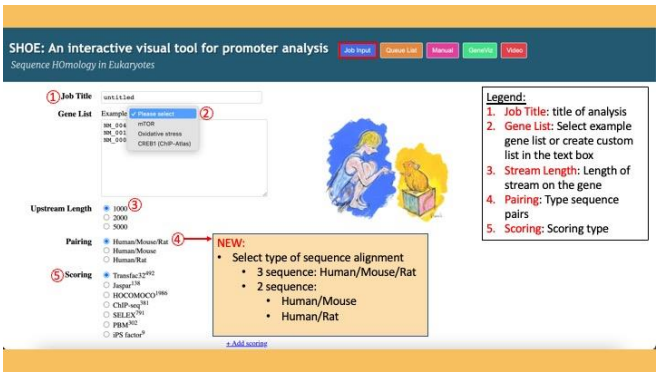
	induction	ns	supression
induction	0	1	91
ns	3	464	3
supression	0	3	1185

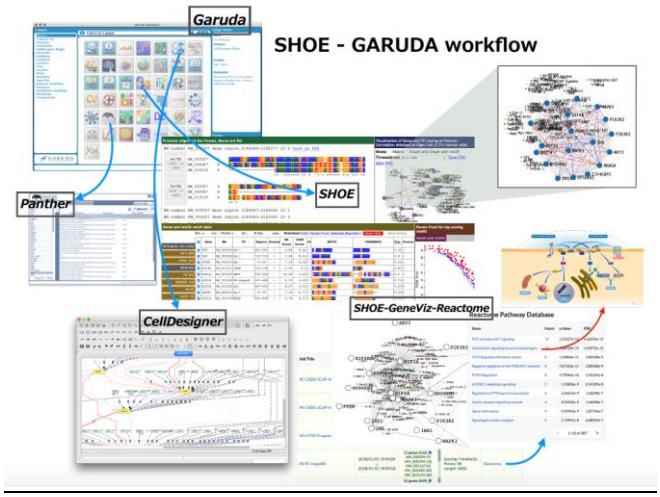
深層学習モデルの予測精度の検証

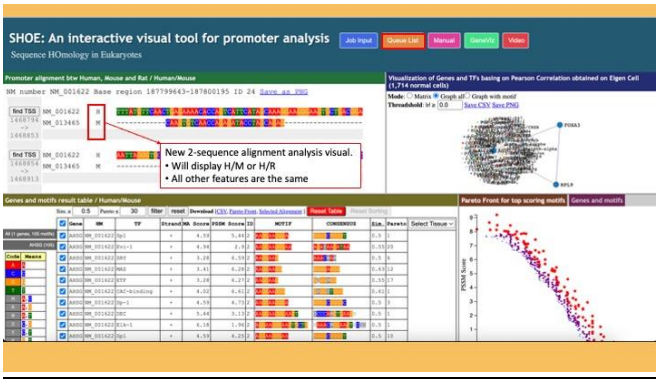
この結果から、induction 群の訓練画像を増幅するのみでは、これらの深層学習モデルは分類に十分な特徴量を抽出できていないことが示唆された。

●転写領域解析ソフトウェア SHOE の改良

今年度は、Garuda Platform 上で、Percellome DB と SHOE の連動をより密にした。下図に連動の概要と、一連の画面を提示する。

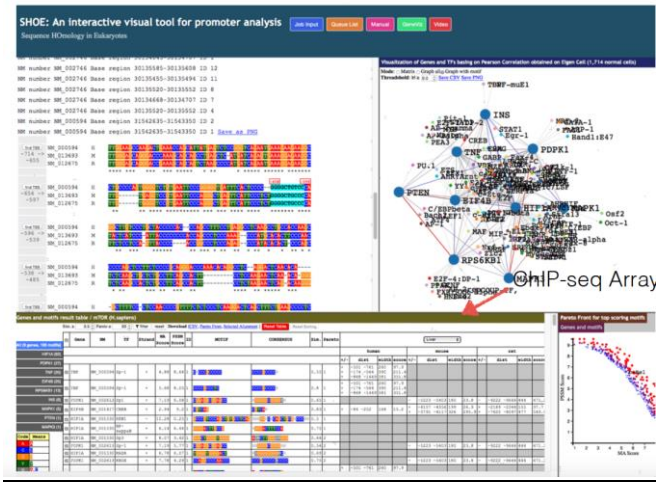


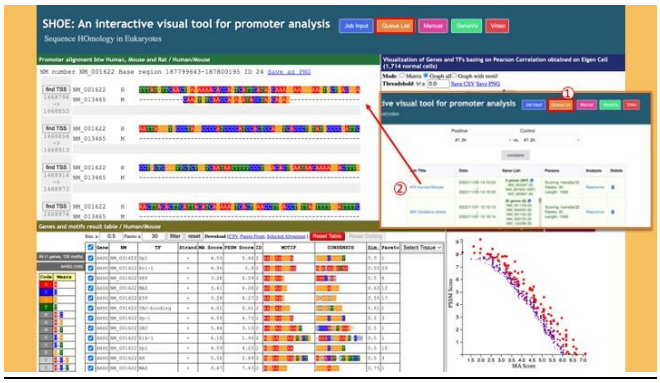


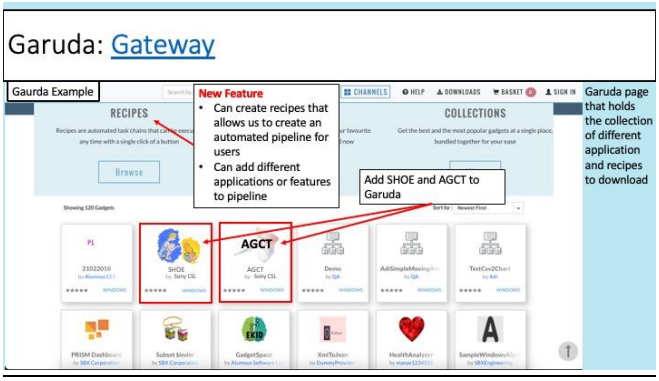


また他の Garuda Gadget との連動を強化し、解析パイプラインの構築を進めた。具体的には Garuda Gateway の上で動作連動を行うレシピ機能を利用して、連続的解析の自動化を行なった。これに合わせて SHOE と ACGT の Garuda 登録を進めた。

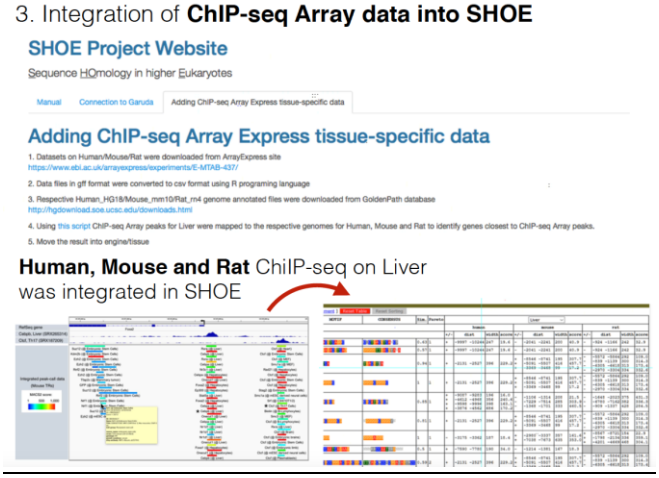
2. Integration of ChIP-seq Array data into SHOE

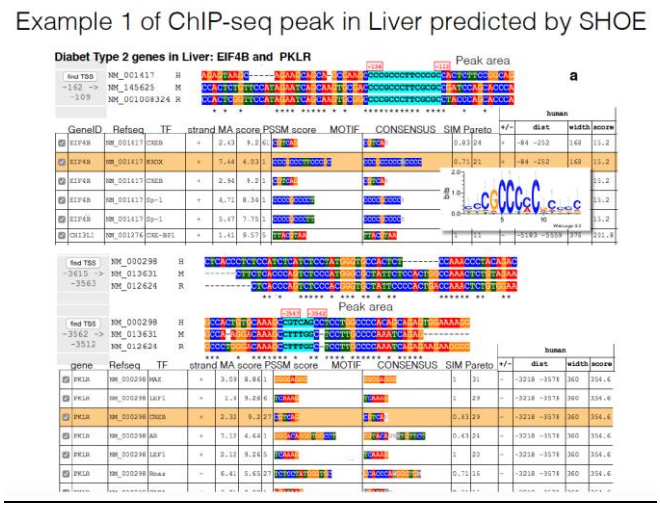


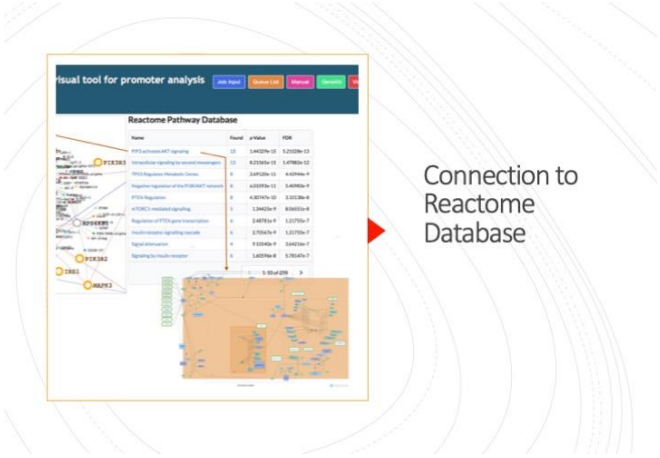


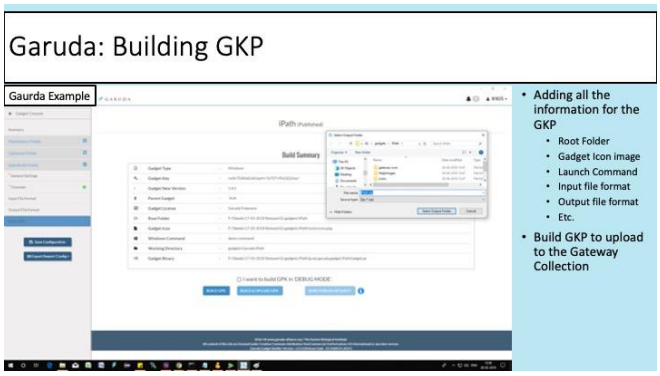


3. Integration of **ChIP-seq Array data into SHOE**









なおクラスター解析に関して今まで使用していた ClustralW がサポート終了のため、ClustalW から Clusteral Omega に変更した。

D．考察

本分担研究においては、独自開発した深層学習遺伝子発現グラフ画像解析システム DTOX においては一般的なバイオインフォマティクス解析パイプラインの精度を大きく上回る性能を達成したことから、深層学習は、大規模データの自動分類に有効であることがわかった。

一方、エピゲノム解析への応用研究については、induction 群の訓練用画像の不足から、一部の分類が困難であり、これを解消すべく使用した生成モデル FastGAN が増幅した induction 群の訓練画像のみでは、深層学習モデルは分類に十分な特徴量を抽出できていないことが示された。

この原因としては、使用している深層学習アーキテクチャは Image Net でプレトレイニングしたものであり、これを転移学習するのみではエピゲノム画像分類に適した特徴量を抽出できていないためであると推察される。そこで、この問題を解決するために、①よりエピゲノム修飾画像に特化したモデルを構築すべく、転移学習ではなく、エピゲノム修飾の訓練画像のみで、一から深層学習モデルをトレイニングする試みと、②より有用な特徴量を抽出することを目的として、言語翻訳ＡＩで活用されている “attention layer”、特に医療画像分類で高い精度を上げている、attention layer の 1 種である、“soft attention”（参考文献 4）の実装と深層学習モデルへの組み込み、を予定している。

E．結論

本分担研究についてはほぼ計画通り推移した。先行研究により開発した遺伝子発現解析用 AI の実装ソフトウェアから、深層学習は、大規模データの自動分類に有効であることがわかったが、新たに開始したエピゲノム解析の AI 自動化の試みからは、充分量の訓練用画像が重要であることは当然として、その構

造に関してはさらに精密なデザインが要求されることが分かった。

一方、解析プロセスの自動化については、初期段階であるが、極めて効率的であることが分かった。

今回の成果で、AI を使った高精度解析とその自動化への可能性が明確になり、課題も同定された。今後は、それらの課題の解決を行う。

(参考文献)

1. Ian J. Goodfellow et al. (2014) Generative Adversarial Networks. https://arxiv.org/abs/1406.2661
2. Vaswani A et al. (2017) Attention Is All You Need. https://arxiv.org/abs/1706.03762
3. Liu B et al. (2021) Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. https://arxiv.org/abs/2101.04775
4. Datta K.S et al. (2021) Soft-Attention Improves Skin Cancer Classification Performance. https://arxiv.org/abs/2105.03358

F．研究発表

１．論文発表

(1) Natalia Polouliakh, Takeshi Hase, Samik Ghosh, Hiroaki Kitano: Toxicity Analysis of Pentachlorophenol Data with a Bioinformatics Tool Set. Methods Mol Biol. 2022;2486:105-125. .[DOI: 10.1007/978-1-0716-2265-0_7]

2. 学会発表

(1) 長谷武志、谷内江綾子、Samik Ghosh、北野宏明 : AI 駆動型オミックスデータ解析とそのシステム毒性学・創薬研究への応用. 第 49 回日本毒性学会学術年会、(2022.7.2)、札幌コンベンションセンター、シンポジウム、口演.

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