MAGICIAN 養成講座

MAterials Genome/Informatics and Chemo-Informatics Activate Networks

第１回イントロダクション

2018.8.22　非常勤講師　山本博志

2015 年 6 月 15 日の日本経済新聞に“米，「材料ゲノム」の衝撃”と言う記事が載った。オーダーメイドの医薬品の開発の際に，患者の遺伝子解析を行って薬を設計する。同じように材料もゲノム解析して設計してしまおうと言う発想だ。筆者は 1997 年頃から，ニューラルネットワークを使った“物性推算と逆設計”を行ってきたので，特に目新しい記事では無かった。しかし，ここに来て様々な団体が人工知能(AI)と材料設計を結びつけるプラットフォームなどを立ち上げているので，“むやみやたらと AI を恐れるな，でも簡単な話なので無視はするな”と言う話を書いておこうと思う。ちなみに，友人の Kevin (Joback 法の物性推算で著名)にゲノム，ゲノムと言っても通じず，スペルを書いたらそれは”ジーノム“と言わなければ通じないと笑われた。

日経新聞の要旨

最新の情報科学を駆使して，優れた性能を持つ新材料の開発速度をこれまでの２倍に高める。そんなプロジェクト，「マテリアル（材料）ゲノム計画」が 2011 年アメリカで始まり，中国でも似たプロジェクトが立ち上がった。材料のデータや論文のデータベースを人工知能が学習して，職人の技の本質を理解して材料を設計する。実験など一度もしなくても，情報科学の手法だけから競争相手の企業と同じ材料にたどり着けた。日本の産業競争力を維持するためにも，危機感を持って取り組むべきだ。という論旨だ。（こうした論旨を捉えるのは AI には難しいらしいが，筆者は何点もらえるだろうか?）

参入の障壁

筆者は化学系の研究者なので身びいきはあるにしても，日本の産業競争力の根幹は素材・材料だと思う。しかし，こうした材料設計は筆者が大学生に教えている程度のことを理解できれば，簡単に行うことができる。学力の差による参入の障壁は大きくない。アイデア勝負のところが大きい。「日本人は英語の論文は読めるが，外国人は日本語の論文は読めない」などと言う障壁も、Google Translate で、今はほとんどなくなってきている。高価なスパコンやソフトが必要なわけでも無く，誰でも気軽に材料設計をすることができる。現在のスマホの計算能力は 20 年前のスパコンより早くなっている。ネットワークも格段に早くなり，情報はネットの上にあふれている。誰でも簡単にマテリアル・ゲノムを始められる環境にある。10 年以上前の資料だが，日本は年 33 万件特許を無料公開し，中国からは 17,000 件/日，韓国からは 55,000 件/日アクセスがあるという。日本も昔はさんざっぱら欧米の特許、論文を読んで、追い付け追い越せしてきたのだから文句は言えない。しかし、単に特許を読んでトレースするのと、AI を使った情報解析の差は認識しておくべきだろう。ノウハウとして隠しておいたことが解析によって明らかになり、逆に特許を握られることも起こり得る。以前なら大企業同士の争いの特許紛争が AI 時代には当然変わってくる。一度も合成したこともない会社が特許の根幹を握ることもあり得る。

AI が学習するにはビッグデータが必要という間違った参入障壁が言われている。ビッグデータをもつ、Google, Amazon, Facebook などがイニシアチブをとり、少ないデータしか持たないなら参入は無理だと言われている。Google は１千万枚の画像を見せてニューラルネットワークに学習させたところ、猫の写真を認識できるようになったという。でも、うちの子供は図鑑 2-3 冊で、猫ぐらい認識できるようになった。自然言語認識、画像認識でビッグデータが必要なのは、AI のレベルが低いだけのことであって、参入障壁にはならない。全てのマテリアルに適用できる汎用化学系 AI ならともかく、特定のマテリアルに限った限定 AI ならデータ数は少なくても十分参入は可能だろう。

大学での授業

横浜国大での授業は今年で 8 年目になる。以前はトピックス（トヨタが燃料電池自動車を売り出した、オリンピックの開催が決まった、など）に合わせて，「我々化学系にできる事」を授業で取り上げて来た。ここ数年は，「AI-ロボティクス時代を乗り切り，後 40 年間職を失わないた

めには今何をしておかなくてはならないか?」を教えている。キーワードは,「AI アシストを受けた化学系研究者」になるためには? だ。 電動アシスト自転車は漕ぐ力の半分をアシストしてくれる。漕ぐ力がゼロならアシスト力もゼロになる。なんでもかんでもAIがやってくれるなら,自分を磨くというモチベーションが無くなってしまうが、化学系ではそうはなりにくい。同じAIを使っても、AIに何を、どう教えたかでAIの答えも変わってくる。結局、地力の高い学生が生き残る。

知識量だけでいえば、とっくの昔にグーグル先生に我々教師の側もかなわない。知識を教えるだけの教師は早晩需要はなくなるだろう。ネットに溢れる知識を全てAIに学ばせたらば、AIの答えは同じになってしまうのではないかと思われるかもしれない。電気・電子のデバイス設計、回路設計のような物理法則に基づくもの、将棋や囲碁のようにルールが明確で、勝ち負けがはっきりしているものは徐々にそうなって行くだろうと思う。しかし、化学は勝ち負けがつかないことが多い。硬いポリマーと柔らかいポリマーのどちらが勝ちですか?と聞かれても「用途次第です」としか答えようがない。そこで、誰が、何を、どう教えたのかによってAIの答えてくれるレベルは千差万別になる。人間としての大事な特徴として、「臨機応変に変化できること」が挙げられる。化学系のAIに対して、教育係はどう教えれば良いかを試行錯誤する。自動化の波が追いつく頃には、それまでに教育したAIを組み合わせて「臨機応変に」次のレベルのAIを組み上げる。「自ら変化し続けられる」ことを売り物にできない学生も(さしあたっては売り手市場であるらしいが)早晩需要はなくなるだろう。
最近更新をサボっているが、大学での講義内容はこちらのページを参照して頂きたい。

https://www.pirika.com/JP/DIY/index.html

化学とボードゲームの違い

AIが人間を超える技術的特異点(シンギュラリティ)を囲碁や将棋の棋士たちは2017年に経験した。後20年や30年は人間が勝つだろうと言われていたのに、あっという間に追い抜かれてしまった。「AIによってなくなる仕事」などが色々言われているが、「仕事を取るのはいいけど、ホビーを取るな!」と言いたいのは筆者だけだろうか?
これまで、囲碁や将棋でAIが勝つのが難しかったのは、ビッグ・データが無かったのがその一因だろう。石や駒を打つ様々な局面に対して、名人だったらどう打つかのデータが圧倒的に足りないので、AIへの教育がうまくいかなかった。転機はある程度、能力が高くなり、後は自己対戦で延々とビッグデータを増やすことができた事だと思う。勝ち負けがはっきりしているのでそれが可能であった。では、ある化合物の毒性(薬効)を予測するシステムを作ることを考えてみよう。毒性(薬効)が既知な化合物はせいぜい数千化合物もない。それをAIに教育しても、圧倒的にデータが足りない。自己対戦させると言っても、どちらの毒性(薬効)予測システムが勝ったのか、負けたのかは実際にその化合物が作られて評価されてないとわからない。コンビ・ケムを使ってハイ・スループットで合成評価する試みが進んでいるが、ボードゲームの自己対戦と比べ、高速化は非常にしにくい。

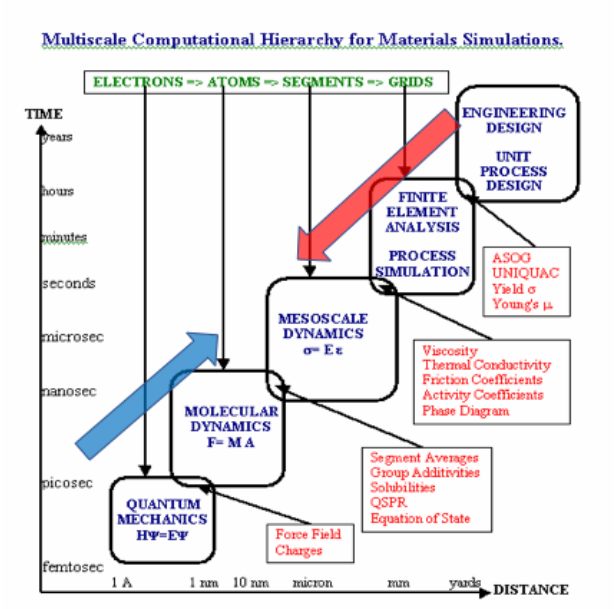
また、ボードゲームは、大きくはチェス、囲碁、将棋、オセロぐらいであるが、化学は、例えばCAS番号が付いている化合物は3000万種、工業的に生産されている化合物10万種、年間1000トン以上生産されているものは5000種と非常にバリエーションが広い。このバリエーションの広さが、一般化の難しさ、ひいては1つ1つの材に対する少ないデータ数になり、AIによる解析が進まない原因となるのだろう。

いわゆる計算機科学との関係

筆者がニューラルネットワーク法(NN法)を始めた1997年の頃から、NN法の問題点として、計算に時間がかかる、メモリーを大量に使う、質の高い学習データを大量に必要とするという問題があった。20年たち、CPUはどんどん高速化し、メモリー、ハードディスクは潤沢になったが、質の高い大量データが必要なのは相変わらず変わっていない。そこで利用されたのが、分子軌道計算(MO計算)や分子動力学(MD計算)などの計算結果などだ。分子軌道計算は、とりあえずは、分子の構造さえあれば計算できる。すると、分子体積、分子表面積、電荷、ダイポールモーメント、HOMO, LUMO, 生成熱などの計算結果が得られる。こうした計算結果を使って“水増しした”ビッグデータをNNの入力に併用することは昔から普通にやっていた。ただ, そうした計算結果を用いたとしても, 例えば分子の沸点や臨界温度を精度よく推算できるようにはならなかった。AIに教えられる化学の知識がまだまだ足りないのだろう。真空中の1分子を、分子集合体にしようとした時に、液体の誘電率与えて計算することがよく行われるが、液体の誘電率を構造のみから予測することすら非常に難しい。ダイポールモーメントの値も、実測値と分子軌道計算値は大きく異なるものも多い。最近流行りのMI(Material Informatics)もベースの部分はMO, MD計算である事が多いようだ。どんなものも計算できるので、

ビッグデータは得やすいが、原子が集まって分子になって、分子が集まってマテリアルになるというほど簡単な話ではない。ただし、計算機科学も利用できるに越したことはない。

下図は、カルフォニア工科大学の Goddard 教授の作成した、Multiscale Computational Hierarchy の図である。大元の図は 30 年以上前に作られたものだが、このバージョンは 20 年ぐらい前のものであろう。Goddard 教授が日本の計算機科学に与えた影響は非常に大きい。日本の計算機科学関連の国家プロジェクトも、この Multiscale Computational Hierarchy を基礎においているものがあるくらいである。



私は、Goddard 教授のところに、1990-1991 年の 1 年 3 ヶ月留学させて頂き、計算機科学を学んだ経験を持つ。高分子合成の実験系の化学者だけどプログラムをつくるのが大好きなので、とても有意義な留学生活を送ることができた。図中の赤い矢印と青い矢印は私が書き加えたものだ。原子、分子、メソスコピック領域へと Goddard 先生は進んで行こうとしておられたのだが、実験系の自分は実際のマテリアルを分割してメソスコピック領域へというコンセプトだった。そうした自分の影響は Goddard 教授も色濃く受け入れてくださっていて、ASOG 法（日大の栃木先生が行なっていた気液平衡の推算法：Analytical Solution of Group）が UNIFAC ではなく記載されている。Activity Coefficients, Group Additivities, Solubility, QSPR などは私が得意とするところとして、図中に記載されている。

ジーノム解析

ジーノム解析とは遺伝子解析の事だ。最新のオーダーメイド医薬の開発では、各人の（とは言っても金持ちだけ？）遺伝子を解析して、個々人の遺伝子的になりやすい病気を特定して予防薬を投与するとか、遺伝子から特定される特異的に良く効く薬を開発するとかが行われている。この発想と同じように材料を遺伝子に見立てて“適者生存の法則”を使って用途に適合した材料を作り出すやり方が、遺伝的アルゴリズム法というやり方だ。別に最近出てきた新しい方法でもなく、2000 年頃から普通に使われているやり方だ。ちなみに、東大の建築科を卒業の女優、菊川怜さんの卒論のテーマは、遺伝的アルゴリズムを使ったコンクリート配合の最適化だったそうだ。この方法を簡単に説明するならば，競馬のサラブレッドをどう育てるかと同じだ。素人考えで恐縮だが，例えば競走馬の強さが，足が強い，心臓が強い（酸素を効率的に血流に乗せて送れる），肺が強い（酸素吸収能力が高い）の３点であったとしよう。強さを表す遺伝子がどれかがわからなくても，実際にレースをして勝った馬だけが子孫を残していけば良い。徐々にレースに勝つ馬が“適者生存の法則”によって遺伝子を残し，生き残っていく。牛乳をいっぱい出す牛，収穫の多く美味しいお米を作る稲など原理は皆同じである。最近面白い話を読んだ。アフリカ象は大きな牙を持つのが特徴なのに、地域によってはメスの 98%が牙を失ったという。これは密漁によって牙を持つ像は淘汰され、牙を持たない象が子孫を残した結果だという。たかだか 100 年ぐらいで適者だけが生き残るので、コンピュータの中で高速に回せばそれなりに素早く解にたどり着ける。

化学の場合の素材に関しては，どういう物性を持つものを“適者”と呼ぶか，それが難しいところでもあり，非常に面白いところでもあり，化学者のセンスが問われるところだ。競馬のレースのように簡単に勝ち負けが決まらないところは，ボードゲームとの違いのようでもある。本来は、遺伝子はビット（0,1 の数列）で表すが、あまり数学にこだわらなくても、それなりに最適値探索は可能であるようである。悩むよりはコードを書いたほうが早いだろう。ここでコードを書くとかいうと、化学系の研究者はちょっと引いてしまうかもしれない。遺伝的アルゴリズムのアルゴリズムとは何かというと、“考え方“と理解すれば良い。あくまでも考え方を示しているだけで、NN 法のように汎用のパッケージソフトがあるわけでは無い。現象ごとに遺伝子の形が異なり、最適と呼ぶ性質も異なる。交叉や突然変異の起こし方も現象ごとに異なる。そこで基本形はあるにしても細部は自分で作り込んでいかなくてはならない。逆にプログラムの専門家であっても、化学のことを分かっ

ていないなら、遺伝的アルゴリズムを正しく動作させることはできない。それにコードを書くと言っても、非常に簡単なコードなので、ぜひトライしてもらえればと思う。
大事なポイントは、遺伝子の定義の仕方と、遺伝子の“適者”の評価の仕方だけだろう。

総当たり法

ガラス、合金や触媒など組成を自由に取れる材料を開発する場合には、遺伝的アルゴリズム法は有効である。しかし、低分子の設計程度あれば、全ての分子をコンピューター上で組んで計算してしまうのも一つの手である。例えばフロン代替化合物の設計では、炭素の基本骨格に対して、水素をフッ素や塩素に変換して全ての構造を自動的に発生させる。そして物性を予測して、欲しい範囲に入った場合に候補化合物として出力させれば良い。1999 年当時、そうした方法でフロン代替化合物をスクリーニングしたところ、当時使っていた 400Mbyte のハードディスクが、炭素数が 7 まで行ったところで溢れてしまい、炭素数は 6 までに制限した。今であれば、オンメモリーでも十分計算できる。
ただし、最近、200 万化合物を計算し、結果をエクセルに貼り付けようとしたら、200 万行はできなかった。ファイルを分割してしまうと、物性値をソートして、良いものを抽出するのは今でも結構大変だった。正確に言うと総当たり法はマテリアル・ジーノムには該当しないかもしれない。しかし、材料設計では未だに重要なポジションだと思っているので、説明に加えたいと思う。

ビッグデータの収集

先にも述べたように、ビッグデータがマテリアル・ジーノムに必須とは思っていないが、最近の自然言語解析、画像解析技術の進歩によってビッグデータも格段に得やすくなった。例えば、ネットに溢れる論文、特許の中から、化合物の名称や分子構造を自動で抽出して、CAS 番号などに紐づけるプロジェクトなどもあると聞く。グラフからデータ点を抽出するくらいは、20 年以上前からソフト・デジタイザーがあったし、大学の授業でもそれを利用して講義を行っている。OCR(光学文字認識)に対して、OSR(光学構造認識)のソフトがアメリカ NIH の HP にあって、PDF などから構造を抽出してくれる。
(https://cactus.nci.nih.gov/cgi-bin/osra/index.cgi)
そうした技術の進歩にもアンテナを高く持つことは忘れないようにしよう。

雑記

化学の世界で、工場に勤務される方が、研究所に勤務する方に言う有名な言葉がある。
「期待はするけど、当てにはしない」
まだまだ、人海戦術で突貫仕事する方が、マテリアル・ジーノムより有効な事は多いだろう。
そう言われた時に考えて欲しい一つの有名な話がある。
「ここに、1 分間で分裂・成長を繰り返すバクテリアがいます。それを 1 つコップに入れたところ、1 時間でコップが一杯になりました。コップの半分に達した時間は幾つでしょう？」
自分の 1/10 の能力しかないと見た時に、自分が 10 倍の能力を獲得するのに掛かった時間で、後 10 年、20 年は大丈夫と思うのが人間であろう。それでも囲碁、将棋はあっという間に追い抜かれてしまった。いくら AI に追い抜かれても、「囲碁将棋ならでは」の楽しみは残るけど、製造業で AI に追い抜かれたら、「実験が趣味だから」と言っても中々認めてもらえないかもしれない。AI アシストを受けて能力を強化した人材になるしか道は残されていないように思える。それに関して最近読んだ新聞記事に面白いことが書いてあった。「あなたは、アトム派？　サイボーグ 009 派？」アトムは人工知能を持ったロボットだが、サイボーグ 009 は能力を強化した人間だ。企業の経営者から見たら、化学系アトムだろうが化学系サイボーグ 009 だろうが成果を出してくれさえすれば良い。大学は化学系サイボーグ 009 の育成を目指さなければ存在意義は無いと思うが、最近の大学は成果直結型の PJ に忙しく、育成はなおざりになっているように感じる。

先日，高校２年生の息子と AI-ロボティクス時代の職について議論した。これからの時代，トップ，ボトムの 15%以外のボリュームゾーン 70%の職は AI-ロボットが行うようになり，人間は労働から解放される時代が来る。そうした時代に必要なのは，ベーシック・インカム（最低収入）で，その収入内で自由に暮らせば良い。そんな記事が新聞に載っていたので，それに関する議論だ。子供に言わせれば，「そんな社会は誰も望んでいない」「自由にと言っても毎日何して暮らせばいいのか」と憤懣やるかたない様子であった。まー，毎日俳句でも考えて暮らすとか（非常に高い確率で俳句の良し悪しを判断するのは AI だろうが），哲学的思索にふけるとか。ただ，原子力もそうだが，「そんな技術は人類を幸せにはしない」と言ったところで，日本だけが止めれば済む問題ではない。日本が完全な鎖国をするか，「人類の幸せ」を真剣に考えてくれる AI を作るか。

星新一の世界になって来る。

それまでは，とりあえずできることをするしかない。筆者がこれまでに授業等で話してきたことをまとめるので、興味のある学生は参考にして頂きたい。計算に必要なソフトや Excel のシートは大学のメールアドレス(XXX@XX.ac.jp)を通じて連絡してもらえればダウンロードのページを案内する予定にしている。

Pirika マテリアル・ゲノム のページ
以下 PDF
第 1 回 イントロダクション 2018.8.23
第 2 回 データ収集と昔ながらのやり方 2018.8.24
第 0 回 物性推算と逆設計と呼んでいた時の話 2000.8.28
なんと 18 年前！
第 3a 回 ポリマー設計と 3 つの MI（その 1） 2018.9.3
第 3b 回 ポリマー設計と 3 つの MI（その 2） 2018.9.3
第 4a 回 MI に適した簡単なデータベースの利用法 2018.9.4
第 4b 回 複雑なポリマーのデータベース化 2018.9.7
プレゼン用：MI を使う時のデータベース構築法 2018.9.11
プレゼン用：複雑なポリマーの設計と DB 2018.9.15
第 5 回 データのクレンジング 2018.8.28
第 6a 回 ニューラルネットワーク法の初歩 2018.9.25
第 6b 回 ニューラルネットワーク法を使った Drug Design 2018.9.22
第 7 回 遺伝的アルゴリズム(GA)を理解しよう