

戦略分野 1: 予測する生命科学・医療および創薬基盤

—生命科学における高性能計算

柳田 敏雄 (理化学研究所 650-0047 神戸市中央区港島南町 2-2-3 e-mail: yanagida@riken.jp)

本戦略分野では、京速コンピュータ「京」の演算能力を用いて初めて可能となる大規模シミュレーションと最先端計測により産出される大量データの計算科学的解析を通して、生命システムの定量的な記述と生命現象の理解と予測を実現し、その予測を創薬や医療などのデザインに応用する。

1. 序

現在、一分子計測技術、質量分析技術、次世代シーケンサーや X 線自由電子レーザー (XFEL) など生命科学に応用可能な先端的な計測手法が次々と開発されており、生命科学の定量化が急速に進展している。これらの計測手法を基盤として、生命システムを物理・化学的な系として量的に記述し、その振る舞いを計算機により予測することが可能になってきた。

本戦略分野では、「京」利用のための高度化推進並びに人材育成、また 4 つの研究開発課題を進める。以下、ここでは 4 つの研究開発課題について紹介する。

2. 細胞内分子ダイナミクスのシミュレーション

京速コンピュータを用いることにより、生体分子のシミュレーションは大きく発展することが期待されている。すでにこの分野は 30 年以上の歴史があるが、従来の計算と

京速コンピュータを用いた計算ではどこが大きく違うのだろうか。言うまでもなく、生体分子は細胞という環境で分子機械として働いている訳であるが、従来のシミュレーションでは演算量を減らすために必要な近似を行っていたため細胞環境を正確にモデルに含めることは困難であった。

我々が、京速コンピュータを用いた大規模分子シミュレーションを実行することによって目指しているのは、細胞という環境で働く生体分子の姿を計算機の中で再現し予測することである。そこでは、細胞内に含まれるタンパク質はもっとダイナミックな構造変化をしているはずであるし、細胞質中に含まれるその他多くの分子と密接な相互作用ネットワークを形成している。また、細胞内には小胞体やミトコンドリアなど生体膜によって隔てられた細胞内小器官が複数存在しており、その生体膜を隔てたイオンや水、アミノ酸などの物質輸送は生体システムの理解に本質的に重要である。また遺伝情報を司る DNA や RNA は核内に存

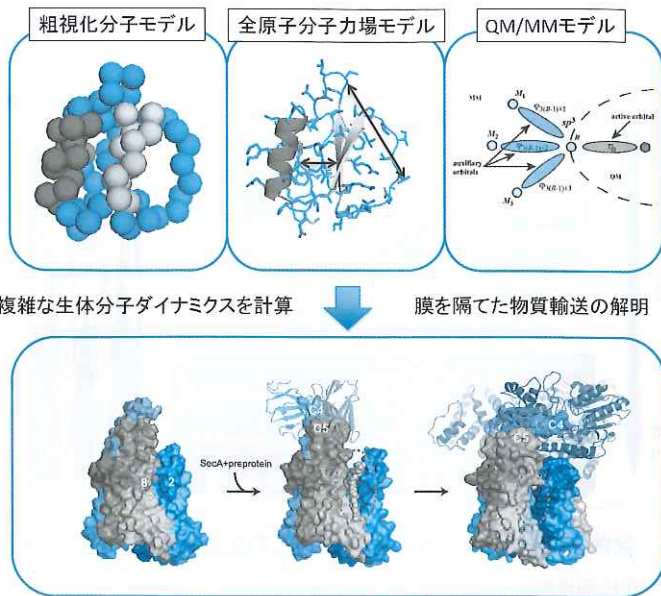


図1 マルチスケール・マルチレゾリューション分子動力学計算法。複数の原子を1個の球で近似した粗視化分子モデル、量子化学計算から近似された経験的分子力場を用いた全原子モデル、活性部位のみの電子状態は露わに取り扱い、その他は全原子モデルを用いるQM/MMモデルを組み合わせることにより、タンパク質の大規模構造変化を計算するための高速で精度の高い分子動力学計算が可能になる。

在しているが、そこでもまたDNAとタンパク質の密接な相互作用による制御機構が存在することが近年の分子細胞生物学によって明らかにされている。このような複雑なシステムをシミュレーションするためには、分子/細胞レベルのシミュレーションにおける多階層/多解像度のシミュレーション技法を連成する必要がある。

そのために平成18年度から理化学研究所を中心に国内の大学や研究機関が連携して「次世代生命体統合シミュレーションの研究開発」に関するプロジェクトが進行しており、分子・細胞スケールにおいても多くの先端的なシミュレーション技法やソフトウェアが開発されている(図1)。また、細胞スケールのシミュレーションを行うために、格子法のみならずタンパク質などの巨大生体分子を1粒子と近似した1分子粒度の細胞シミュレーションなども開発された。

戦略的に細胞内で働く生体分子のシミュレーションを行うために、我々は以下の3つのサブ課題を掲げている。それは、(A)細胞内信号伝達経路のモデリング、(B)膜タンパク質による細胞膜を隔てた物質輸送、(C)核内DNAタンパク質の構造予測と機能解明である。

まず、サブ課題Aでは真核細胞の信号伝達系が行う複雑な情報処理がどのように実現されているかを1分子粒度の解像度で計算する。従来の計算で演算量を減らすために用いられてきた近似(個々の信号伝達経路毎の計算、細胞内環境を考慮しない均一系を想定)から、「京」を用いたシミュレーションでは、信号伝達全体の統合計算や粗視化分子動力学計算と連携した計算などを行い、より細胞に近い環境での信号伝達をモデリングする。その結果、細胞の分化、

増殖、がん化などに関わる細胞の情報処理機構に関する新しい見解が得られると期待される。

サブ課題Bでは膜輸送タンパク質(ポンプやトランスポーター)などによる細胞膜を隔てた物質輸送を解析する。特に、ポンプなどの能動輸送はプロトンやATP加水分解などの外部からのエネルギーを必要とし、膜タンパク質の大規模構造変化を伴う場合が多い³⁾。そのため、輸送速度はミリ秒以上の遅い現象であり、分子シミュレーションでその輸送サイクル全体を解析することは困難であった。この課題では、粗視化/全原子/QM/MMなどを連成したマルチスケール分子シミュレーション法を用いて物質輸送全体の解明を目指す。本研究課題で行うシミュレーションと1分子計測技術などによる実験情報を組み合わせることで、細胞膜内で膜輸送タンパク質が働く姿を分子レベルの解像度で可視化することが可能になる。

サブ課題Cでは核内DNAタンパク質の構造予測と機能解明を行う。全長2mのヒトゲノムDNAが各細胞内に格納されているが、転写される場合にはその活性部位のみがほどける。しかし、染色体がどのような構造を実際に行っているのか、がん化や幹細胞化、老化と関連している塩基の修飾(エピジェネティクス)の構造基盤などにはいろいろな論争が未だに存在する。分子シミュレーションでは現在、最小の単位であるモノヌクレオソームの安定性解析などがようやく可能になったレベルであるが、「京」によるシミュレーションでヌクレオソーム多量体の構造モデリングやエピジェネティクスの解明を目指している。遺伝子高次構造からがんや細胞の老化などの生命現象に関する分子機構が解明される可能性も高い。

3. 創薬応用シミュレーション

生物の体において外界や他の器官からのシグナルを受け取る機能を持ったタンパク質を受容体と呼び、それに結合する分子をリガンドと呼ぶ。受容体によって、低分子化合物、ホルモン、ペプチド、他のタンパク質など種々の分子がリガンドになり得る。受容体とリガンドは一般に化学結合を伴わない極めて弱い相互作用の集積で結合する。あるリガンドが受容体に強く結合すると、そのタンパク質の機能をリガンドが阻害または促進して薬効がよく現れる可能性が高い。このため、種々のリガンドが受容体に結合するエネルギーを正確に求めることが、計算機を使った薬設計(Computer Aided Drug Design: CADD)を実現するための必須条件になる。一般に受容体とリガンドの結合エネルギーは-10 kcal/mol程度であり、薬候補化合物間の優劣を判断するには1 kcal/mol (43.3 meV)の精度で結合自由エネルギーを求める必要がある。

二つの熱平衡状態間の自由エネルギー差を求める様々な計算方法が今までに考案されているが、その中で最も注目されるのは1997年にJarzynskiが発見した非平衡仕事量の等式 $\Delta G = -k_B T \ln \langle \exp(-W/k_B T) \rangle_{AV}$ である。 W は二つの

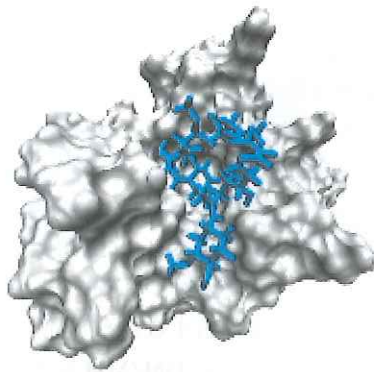


図2 免疫抑制剤タクロリムス(開発コード: FK506)がそのターゲットタンパク質FKBPに結合している構造. FK506は1984年藤沢薬品(現アステラス製薬)が筑波山の土壌細菌から発見した.

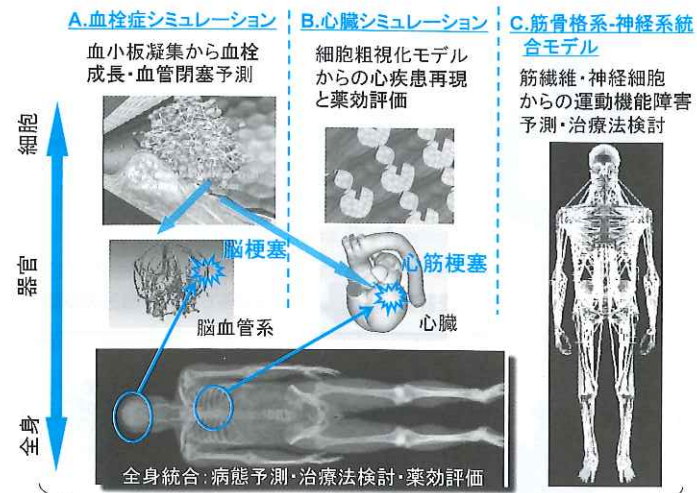
状態間の非平衡仕事量, k_B はボルツマン定数, T は温度で, $\langle \exp(-W/k_B T) \rangle_{AV}$ は仕事量の平均を意味する. 2006年にはCuendetが能勢-Hooverの温度制御を用いた動力学方程式から熱力学極限を援用せずにJarzynski等式が導出できることを示した.⁵⁾ これは限られた原子数で行う分子動力学シミュレーションで自由エネルギーを厳密に計算できることを意味する.

Jarzynski等式を用いて結合自由エネルギーを計算した例として, 臓器移植の時に拒絶反応を抑制する免疫抑制剤として有名なFK506とその標的タンパク質の系(図2)がある. 共結晶構造からスタートして約800個の分子動力学を異なるパラメータで2ナノ秒ほど並列に走らせて, 計算全体でマイクロ秒オーダーの動力学シミュレーションから非平衡仕事量を算出してJarzynski等式で自由エネルギーを求めた.⁶⁾ この例から水分子を陽に入れた高精度力場を使用することで, タンパク質とリガンドの結合自由エネルギーがCADDによる薬設計に必要な精度で求まることが明らかになった.

次世代スーパーコンピュータ「京」を用いると創薬対象となるタンパク質とリガンドが水分子に囲まれた状態で, 数十マイクロ秒の分子動力学を数千個並列に実行できる. リガンドとタンパク質が離れた状態から結合するまでの時間を隈無くサンプリングすることが可能になるので, 実験で結合構造が得られていなくとも結合自由エネルギーを精度良く求めることが可能になる. これによりコンピュータで薬設計を行う必須条件が整うことになる.

4. 予測医療に向けた階層統合シミュレーション

来るべき高齢化社会に向け, 年齢に関係なく人々が健康的にはつらつと生活し, 活気のある社会を作っていくことが期待されている. そのため, 重篤な疾患に対する高度な治療法の確立だけでなく, 健康的な生活が維持できなくなるような疾患に至る兆候を事前に予測し予防する, あるいはできるだけ早く発見し, 患者の身体的負担の極力少ない治療を施すことが益々重要となってきている. 本課題では,



次世代スパコンにより細胞の集合体としての人体の様々な機能の再現

図3 血球細胞・心筋細胞・筋繊維・神経細胞レベルからの全身統合シミュレーションによる病態の再現と治療法の検討, 薬効の評価.

次世代計算機の持つ圧倒的な計算能力を有効に活用し, 細胞レベルに始まる現象の積み上げとして組織・器官の挙動を予測するシミュレーションを実行し, わずかな兆候からの将来の病態の予測や, 身体的負荷の少ない治療法の検討, 薬効の評価などのためのシミュレーションを行っていく.

本課題では, 深刻な後遺症をもたらし得る心疾患・脳血管疾患および, 運動機能障害をもたらす神経疾患を対象とし, これまで別々に開発を進めてきた血栓症シミュレータ, 心臓シミュレータ, 筋骨格シミュレータ, 脳神経系シミュレータの統合により, それぞれの疾患に対する血球細胞・心筋細胞・筋繊維・神経細胞レベルからの複雑なプロセスを取り込み, 病体の予測と治療支援, 薬効の評価などを行う. 具体的には, サブテーマとして(A)心筋梗塞・脳梗塞の直接的な原因となる血栓による血管閉塞を再現するマルチスケール血栓シミュレーション, (B)心筋細胞レベルから心臓全体まで様々な心疾患とその治療法・薬効の評価も可能なマルチスケール・マルチフィジックス心臓シミュレーション, (C)筋肉・臓器の体積効果, 筋膜・腱膜の表面効果を取り入れた全身筋骨格-神経系統統合シミュレーションを実行していく(図3).

サブテーマAでは, 心筋梗塞へ至るプロセスを再現するため, サブテーマBで開発されている心臓シミュレータとの連成を行う. これにより, 成長した血栓が血管を閉塞することによる心筋梗塞, すなわち, 心臓をとりまく血管である冠状動脈において, 動脈硬化により狭窄した血管壁に血小板が吸着し, 血栓へと成長したのち血管閉塞に至るプロセスを再現する. 血小板と血管壁との相互作用に関する分子論的解析から血小板全体と血管壁との間の接着力を評価し, 血栓を引き剥がす流体力学的な力と併せて, 血栓成長から血管閉塞への詳細なプロセスを再現する. 閉塞する血管の大きさが数mm程度, 血小板の大きさが $2\mu\text{m}$ 程度

であり、次世代スパコンにより初めて、血小板を十分解像し、かつ実形状に合わせた3次元計算が可能となる。

サブテーマBでは、すでに開発済みの心筋細胞内の各種イオン電流や収縮タンパクの挙動を表す生理学モデルから、心筋組織の収縮・弛緩、そして心臓全体の拍動と血液の拍出、さらには血圧・心電図までを再現するマルチスケール・マルチフィジックス心臓シミュレータ⁸⁾を用いる。そして、次世代スパコンの10ペタフロップス級の計算機の機能を最大限に利用し、様々な条件下における種々の疾患(不整脈や拡張型心筋症など)の再現、薬効の評価を含む種々の治療法を検討するためのシミュレーションを行っていく。また、本課題では、循環器系-神経系-筋骨格系の全身階層統合シミュレーションに向け、神経系の影響として、交感神経系・副交感神経系が心臓の活動に与える影響についてモデル化を行い、既存の心臓シミュレータに取り込むことを検討していく。

サブテーマCでは、骨の剛体多面体モデルからなるリンク機構と1次元の筋肉機能モデルをつなぎ合わせたヒト筋骨格系全身モデル(リンク数53本、筋989本)に末梢神経系のネットワークを取り込んだモデル⁹⁾をベースにし、さらに筋肉・骨・腱・関節などと皮膚や臓器の3次元構造を組み込み、全体を連成させた大規模計算を行う。また、人の運動機能や心臓をはじめとした循環器系の機能に大きな影響を与える神経系の役割について神経細胞レベルからの物質輸送・シグナル伝達の影響を考慮した解析を進める。神経系が人の運動に与える多岐にわたる影響に対して、筋骨格系モデルと脳神経系モデルさらに認知活動を含む行動モデルの統合により、シミュレーションを実施し、神経疾患による運動機能障害の再現やリハビリ方法の検討などを行っていく。さらに、神経系のモデリングを軸として、循環器系・筋骨格系の全身統合を達成し、運動時・興奮時・疲労時等において人体内で起こる様々な変化を予測するための基盤ツールの整備も進めていく。

5. 大規模生命データ解析

この戦略課題は、ゲノムを基軸とした大規模生命データ解析により生命プログラムとその多様性を理解することを目標としている。ゲノムには生命システムを構成する部品的设计図とそれらの部品を組み立てて機能を作り出すおおよそのプログラムが記述されている。細胞から地球規模のシステムを構成している基本情報がゲノムである。

このゲノムを基軸とした生命データを京コンピュータで大規模に解析し、個人個人のがんなどの細胞の個性を生命システムとして描出する研究、予測能力を持ったモデルで病態や生命システムを御す研究、そして、病気や生物多様性を理解するなかで生命科学を再構築する研究などを展開する(図4)。その波及効果は、薬効・副作用の予測、薬の毒性原因の推定、新たな診断技術、標的分子の探索、オーダーメイド投薬、ゲノムによる環境解析技術、ゲノム情報

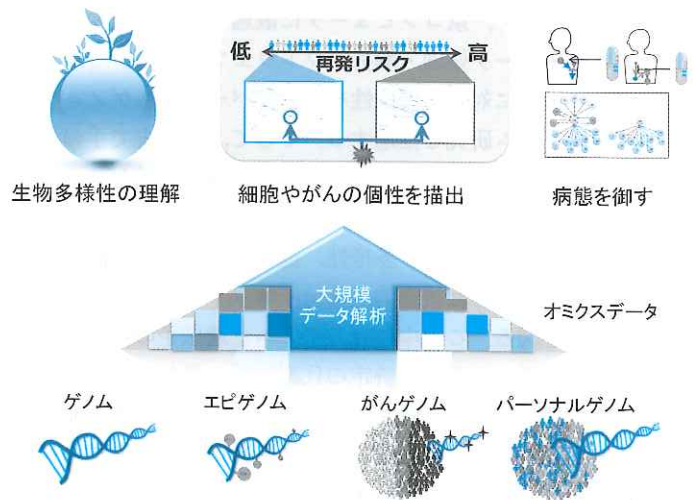


図4 最先端シーケンサーデータの処理・解析ソフトウェアを整備し、ゲノムを基軸とした生命データの大規模解析により生命プログラムとその多様性を理解する。オミクスデータとは、転写産物、タンパク質、代謝物質などの生体内分子の状態を網羅的に計測して得られたデータ。エピゲノムとは、遺伝子そのものには影響しないが、遺伝子のはたらきを変化させるDNAの化学的修飾のこと。

の産業利用など、広い分野に及ぶ。

これまで、生命プログラムの研究は、モデル生物やゲノム地図を頼りに、限られた生命データを小規模に解析することで進められてきた。その結果、要素還元論的な生命システムや病気の理解や、ゲノム変異と病気の統計的関連付けによる原因遺伝子の発見などがあつた。しかし、数年前に登場した次世代シーケンサーおよび、今後登場するシリコンシーケンサーなどにより大変革が起きている。パーソナルゲノム時代が近づいている。パーソナルゲノムを決めるコストが10年以内に1万円を切る時代に入るといわれている。主要ながんのゲノム異常カタログを作成する国際がんゲノム計画は、5万人分の全ゲノム情報を解析する¹¹⁾。また、脊椎動物1万種のゲノムを決めるGenome 10K Projectや、土壌や海中の微生物や腸内細菌をまとめて解析するメタゲノム研究も進んでいる。今後、地球環境を最も直接的に捉えることができるのはメタゲノム解析になるかもしれない。

最先端シーケンサーは、ゲノムのDNA配列だけでなく、DNAの修飾状態、DNAのコピー数、機能性RNAを含む転写産物を量的に一挙に解析できる。プロテオームやメタボロームの解析も精度・速度ともに向上している。さらに、細胞内の1分子動態を捉える計測技術などが開発され、生命システムデータの精緻化・大規模化が急速に進んでいる。こうした状況のもと、生命科学を次世代生命科学へと質的に転換させる一つの戦略は、京コンピュータの圧倒的な計算リソースの下で、最先端シーケンサーデータの解析を推進力とし、生命科学に係る計算科学とインフォマティクスにより生命科学を再構築することである。この戦略課題は、大規模データ解析という方法によりこの変革にチャレンジする。

そこで、まず、京コンピュータに最適化した最先端シークエンサーデータ処理・解析ソフトウェアを整備する。これにより世界に対して優位性を持ってパーソナルゲノム・地球規模ゲノム研究を推進する。そして、これまで手の着けようがなかった大規模ながんゲノム解析、比較ゲノム解析、メタゲノム解析などの研究を「京」により実現し、生命プログラムの複雑性、多様性、進化を理解する。同時に、ゲノムを基軸とした統合的生体分子ネットワーク解析研究に新展開を図る。この中には、ゲノムと遺伝子の転写産物やタンパク質の発現量、特定の代謝物質の量などを形質としたオミックス量的形質を用いた関連解析、薬剤応答ネットワーク解析、そして気象学で威力を発揮しているデータ同化技術を多様で複雑な分子ネットワークモデリングに応用する研究などが含まれる。これにより、がんのシステムとしての個性をはじめとして、生体内分子、化合物、環境因子などが織りなす病態・生命プログラムおよびその多様性を明らかにすることが可能になる。例えば、図4の上中の「個人個人のがんを分子のシステムとして描出」することは、「再発」といったただ一つの尺度に対してだけでも、全遺伝子を対象とした1,000人分の遺伝子発現データを解析するには10テラフロップスの計算リソースを使って約3カ月の時間を要している。「京」の力があれば、サンプル数を増やすこともでき、より精緻に、個人個人のがんの個性を一日のうちに捉えることができるようになる。

6. 終わりに

生命は、複雑・非線形・動的な系の代表である。現在、生命科学においては大規模・精密な定量化が進行しつつあり、シミュレーションの適用対象は拡大しつつある。生命科学・工学分野は、今後の高性能計算の応用において、最も大きな可能性を秘めているといってもよく、同分野の発展のため、本戦略研究を進めていく所存である。

執筆にあたり、全体的に大阪大学の中村春木氏、理化学研究所の木寺詔紀氏、2章は理化学研究所の杉田有治氏、3章は東京大学の藤谷秀章氏、4章は東京大学の高木周氏、5章は東京大学の宮野悟氏より、多大な貢献をいただいた。ここに深謝の意を表する。

参考文献

- 1) J. A. McCammon, B. R. Gelin and M. Karplus: *Nature* **267** (1977) 585—Dynamics of Folded Proteins.
- 2) B. L. de Groot and H. Grubmüller: *Science* **294** (2001) 2353—Water Permeation across Biological Membranes.
- 3) C. Toyoshima, H. Nomura and Y. Sugita: *FEBS Lett.* **555** (2003) 106—Structural basis of ion pumping by Ca^{2+} -ATPase of sarcoplasmic reticulum.
- 4) C. Jarzynski: *Phys. Rev. Lett.* **78** (1997) 2690—Nonequilibrium Equality for Free Energy Differences.
- 5) M. Cuendet: *Phys. Rev. Lett.* **96** (2006) 120602—Statistical Mechanical Derivation of Jarzynski's Identity for Thermostated Non-Hamiltonian Dynamics.
- 6) H. Fujitani, Y. Tanida and Azuma Matsuura: *Phys. Rev. E* **79** (2009) 021914—Massively Parallel Computation of Absolute Binding Free Energy with Well-Equilibrated States.
- 7) K. Sugiyama, S. Ii, S. Takeuchi, S. Takagi and Y. Matsumoto: *J. Comput. Phys.* **230** (2011) 596—A full Eulerian finite difference approach for solving fluid-structure coupling problems.
- 8) <http://www.sml.k.u-tokyo.ac.jp/index.html>
- 9) Y. Nakamura, K. Yamane and A. Murai: *Ann. Int. Conf. IEEE Eng. Med. and Biol. Soc.* (2006) 99—Macroscopic Modeling and Identification of the Human Neuromuscular Network.
- 10) *Nature* **467** (2010) 1026—Human genome: Genomes by the thousand.
- 11) *Nature* **464** (2010) 993—International network of cancer genome projects.

著者紹介



柳田敏雄氏：専門は生物物理学。生体分子の1分子計測、生体分子機械の動作原理、脳記憶のダイナミズムに関する研究など。

(2011年1月19日原稿受付)

Computational Life Science and Application in Drug Discovery and Medical Development

Toshio Yanagida

abstract: In these strategic fields, the K computer project will bring to a realization of quantitative descriptions of biological systems and deliver comprehension and predictions of biological phenomena through a computer science-based analysis of the massive amounts of data generated by state-of-the-art measurements and large-scale simulations that will first be made possible using its computational power. These predictions will then be applied to drug discovery and medical design.