

REVIEW

Integrating statistical inference with mathematical deduction in protein simulation

Diego Liberati^{*†}

National Research Council of Italy, Politecnico di Milano University, Milan, Italy

***Correspondence:**Diego Liberati,
Diego.liberati@cnr.it**†ORCID:**Diego Liberati
0000-0002-4294-6705**Received:** 24 October 2025; **Accepted:** 05 November 2025; **Published:** 16 January 2026

Mathematical deduction and statistical inference are often seen as opposite, competing approaches in order to achieve results, also in protein simulation. Our claim is that, once more, cooperation is, as often, better than competition. A few approaches in this sense are recalled, mainly developed with our pupils over the recent decades, letting us understanding their powerful combination also in protein simulation thanks to a few recalled examples. The mathematical modeling approach is mainly related to piecewise affine identification within the frame of the dynamic-logic hybrid systems, while the statistical inference is within the frame of explainable artificial intelligence with special emphasis on machine learning of rules intelligible to humans.

Keywords: inference, deduction, simulation, schroedinger equations, explainable AI

Introduction

Protein simulation is of paramount importance in both understanding nature and trying to control it by driving or repairing it toward our goals. Such two aforementioned sides, the scientific and the technologic ones, are not even disjoint, one needing to understand before even being able to dare to control, even sometimes by mimicking, at least to a certain amount, what we have learned from nature itself.

In such a contest, a minireview will be provided here trying to put together salient aspects of two pairs of couples of approaches traditionally separated. On one side, knowledge of a powerful modelling task like Schroedinger equations has been pushing over years to search for improved calculus ability to increase the overwhelming number of atoms involved and try to compute every single detail everywhere within the molecule. The course of dimensionality, especially in such a nonlinear complex system, has shown that the limit to such a “not-so-brute-because-bright” but still quite “brute-force” approach is not just increasing calculus need but the increasing of local minima often hiding the very

results, when complexity of the molecule and lack of complementary information let us uncertain in a richness not easily allowing chemo-physical grasp of the whole underlying chemo-information.

It is then quite useful to resort *also* to (not just instead of) the dual coarse approach, limiting the detail to what is really needed to understand, without going into unnecessary details. For this purpose, statistical inference from data could be very helpful at a very little cost, paying attention that there is direct total correlation within statistical and coarse on one side and deterministic and atomic on the other, not that there is any need of the same granularity overall for all the molecules, multiresolution allowing to investigate all details around catalytic and allosteric sites when just describing what is needed elsewhere.

Finally, simulation could be of help to forecast new properties of the in silico designed molecule.

The increased complexity of the approach, like the one analyzed in a different domain by Salami and Khajehvand (1), is worth facing because of the promising results.

Methodology and results: new approaches to complement traditional protein simulation

Understanding nature

In nature, the molecule does exhibit, besides an obvious chemical functional behavior, a kind of mechanical behavior in its 4-dimensional spatio-temporal deployment that is ancillary to the chemical task.

Son of sevenless oncosuppression of Ras

As a matter of example, for instance in SoS1 the mechanical dance of the few domains does determine the amount of exposition of the catalytic and allosteric sites to Ras reagent according to its concentration under the cell membrane: in Sacco et al. (2) a just coarse statistical model of the dynamics of a quite key natural molecule, SoS, fully described in its oncologic properties by just dynamical interplay of its few domains in presence of its target Ras, has been shown to be even able to predict a then discovered mutant, just playing through Gillespie simulation over the Markov model of its statistical step-by-step dynamics, whose parameters have been identified through more standard chemical investigation including plasmon one.

Tumor Necrosis Factor alpha in bystander effect

Analogously, tumor necrosis factor, for instance, released in the so-called bystander effect after radiotherapy, does trimerize before entering the membrane of the cells it signals to either apoptosis or reinforcement (3) should the extracellular concentration be high enough, because only when trimerized could it force the membrane pore, thus transforming a chemical property into a mechanical one, limiting the access to the target, then to a chemical one again, fostering either the apoptotic or the reinforcement path, depending on the inner state of the targeted cell (3).

Mimicking nature

The two above examples are paradigmatic about the power of even such simple approaches, but of course not granting replication on every synthetic molecule or protein: the fact that SoS1 and TNF do really exist, and cooperate in keeping our health, allows to forecast that a useful functional effect is in some way embedded in their chemical structure, not being the same a priori granted for every synthetic compound, not even proteins, whose building blocks are in fact belonging to a very small alphabet of less than a couple of dozen different kind of bricks: if we think about English the situation is similar, just twenty-some letter within the

alphabet are well able to deploy an almost infinite possibility of says, even non sense.

Logical networks for machine learning

Care has then to be put not just in carefully choosing components, but mainly in combining them in such a way that their nonlinear interaction would exhibit the most desired properties while removing the most undesired ones. Of course Schroedinger equations would help with the burden we recalled, but a more coarse statistical approach could at least partly, and even better, help, allowing by examples to compose domains instead of just single atoms, like the above-described approach to SoS1. In this sense, better than popular neural networks, great in learning but poor in explaining, logical networks (4) could be of real help, able to straightforwardly identify the logical function embedding classification or regression on data in the canonical form of OR of ANDs, thus yielding to express for every stationary shot the logical interactions in a deductive language of the predicative kind.

IF (... AND ... AND ...) OR (... AND ...) THEN ...

then immediately understandable in predicative logic, and a posteriori integrable with deductive rules derived from the mathematical model approach, thus in principle letting us exploit goods from both approaches, cooperating instead of competing.

Piecewise affine dynamics

If data are dynamic, the time course could help being directional: this could be simply exploited by identifying the polytope behavior by a piecewise affine dynamic-logic approach (5), reducing the multivariable input on the desired output to linear hyperplanes each confined in also identified boundaries in time, and then allowing even Principal Component Analysis to enhance salient influencers to the desired output.

Principal component analysis to principal directions divisive partitioning + K-means smart clustering

In a complementary setting, genomic in nature, deputed to recall how to build useful proteins, such a simple composition of known approaches has been shown both to outperform (6) a quite known paper on Science (7) and to confirm a just-hypothesized pathway in leukemia differential analysis (8) toward such personalized precision medicine nowadays finally looked for.

Discussions

A few approaches already successful in some contexts have been recalled and proposed in order to be integrated also in drug discovery in silico. Preliminary results, which are still not possible to be disclosed due to a pending patent, are encouraging. Further investigation is, of course, needed to assess if the proposed approach, already successful in other problems in some way quite similar to the one under analysis, could really be of help, maybe together with the other recalled methods.

Conclusion

A portfolio of valuable instruments has been proposed that could also be the bone of a roadmap that could directly be used in understanding and discriminating protein complexity, helping to design synthetic proteins by learning their properties from analogous once while taking into account known priors and Schroedinger computation when available: cooperation (among methods and among scientists from complementary disciplines) is once more the key to success.

Funding

Research was not funded.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Salami Y, Khajehvand V. LSKE: lightweight secure key exchange scheme in fog federation. *Complexity*. (2021) 7:1–9.
2. Sacco E, Farina M, Greco C, Lamperti S, Busti S, DeGioia L, et al. Regulation of hSos1 activity is a system-level property generated by its multi-domain structure. *Biotechnol Adv*. (2012) 30(1):154–68.
3. Mezza D, Pappalettera M, Liberati D. A quantitative numerical model for TNF- α mediated cellular Apoptosis. *Biol Eng Med*. (2016) 1(1):1–8.
4. Muselli M, Liberati D. Binary rule generation via Hamming Clustering. *IEEE Trans Know Data Eng*. (2002) 14(6):1258–68.
5. Ferrari-Trecate G, Muselli M, Liberati D, Morari M. A clustering technique for the identification of piecewise affine systems. *Automatica*. (2003) 39(2):205–17.
6. Garatti S, Bittanti S, Liberati D, Maffezzoli A. An unsupervised clustering approach for leukaemia classification based on DNA micro-arrays data. *Intell Data Anal*. (2007) 11(2):175–88.
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression. *Science*. (1999) 286:531–7.
8. Grassi S, Palumbo S, Mariotti V, Liberati D, Guerrini F, Ciabatti E, et al. The WNT pathway is relevant for the BCR-ABL-1 independent resistance in chronic myeloid leukemia. *Front Oncol*. (2019) 9:532.