Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

Structures, dynamics, complexes, and functions: From classic computation to artificial intelligence

Elena Frasnetti[1,a], Andrea Magni[1,a], Matteo Castelli[1], Stefano A. Serapian[1], Elisabetta Moroni[2] and Giorgio Colombo[1]

Abstract
Computational approaches can provide highly detailed insight into the molecular recognition processes that underlie drug binding, the assembly of protein complexes, and the regulation of biological functional processes. Classical simulation methods can bridge a wide range of length- and time-scales typically involved in such processes. Lately, automated learning and artificial intelligence methods have shown the potential to expand the reach of physics-based approaches, ushering in the possibility to model and even design complex protein architectures. The synergy between atomistic simulations and AI methods is an emerging frontier with a huge potential for advances in structural biology. Herein, we explore various examples and frameworks for these approaches, providing select instances and applications that illustrate their impact on fundamental biomolecular problems.

Addresses
[1] Department of Chemistry, University of Pavia, via Taramelli 12, 27100 Pavia, Italy
[2] SCITEC-CNR, via Mario Bianco 9, 20131 Milano, Italy

Corresponding author: Colombo, Giorgio (g.colombo@unipv.it)
𝕏 (Colombo G.)
[a] These authors contributed equally to this work.

Current Opinion in Structural Biology 2024, **87**:102835

This review comes from a themed issue on **Biophysical Methods (2024)**

Edited by **Gabriela Chiosis** and **Elizabeth A. Komives**

For complete overview of the section, please refer the article collection - Biophysical Methods (2024)

Available online 13 May 2024

https://doi.org/10.1016/j.sbi.2024.102835

Keywords
Molecular simulations, Molecular dynamics, Biological complexes, Machine learning, AI, Drug design.

Introduction

Recent developments in the investigation of biological mechanisms have substantially changed our view of how proteins work: the dominant model has gradually shifted from the reductionist view in which one protein sequence corresponds to one structure and one function, to a new view in which all proteins are dynamic entities that sample distinct structural states and engage in different complexes with other biomolecules [1–3].

The dynamic nature of proteins and their ability to collaborate with different partners depending on the specific environment or cell needs is what allows context-specific function to emerge. Paradigmatic examples are the ribosome, the proteasome, or the chaperone machinery, in which the components select the most-suitable conformational states for interactions and assemble to achieve their functional goals. In this context, there is growing evidence that the structural forms of interacting proteins and the dynamics of interconversion among them can be further fine-tuned by the impact of post-translational modifications, which may be different in health and disease.

Understanding how these variations influence context-dependent functions, shedding light on the mechanisms that underlie interactions, and developing methods to (re)design complex assemblies can significantly impact our understanding of chemical biology and the way we use this knowledge to develop chemical tools and therapeutics.

Dramatic advances in experimental approaches, ranging from transcriptomics to proteomics and structural resolution techniques, are providing an unprecedented level of information on complex protein organizations [4]. At the same time, novel ways to screen for biologically active ligands and to develop protein-based interactors (such as, for instance, therapeutic antibodies) are expanding the arsenal of molecules that can be exploited for basic research and therapeutic purposes [5,6].

In parallel, new approaches based on integrative modeling [7] have been gaining prominent momentum over the last few years, providing a high resolution picture of large and complex multiprotein assemblies. In this context, Mosalaganti et al. [8] developed a model for the structure and dynamics of the human nuclear pore complex combining AI, cryo-electron tomography (Cryo-ET) and coarse-grained MD simulations. Singh et al. [9] reported the use of in-cell cryo—electron tomography and subtomogram analysis to investigate the cage-like nuclear basket, a peripheral region of the nuclear pore complex which shows significant variation among species. Using integrative modeling, the authors computed a model of the basket in yeast and mammals that revealed how a hub of Nups in the nuclear ring binds to basket-forming Mlp/Tpr proteins, forming a docking platform for mRNA recognition and preprocessing before nucleocytoplasmic transport. A notable example of the combination of enhanced sampling molecular dynamics (MD) simulations with adaptive Markov state modeling, cryo—electron microscopy (cryo-EM), small-angle x-ray scattering, and hydrogen—deuterium exchange mass spectrometry has been reported by Juyoux et al. [10] to describe the structure and dynamics of mitogen activated protein (MAP) kinase phosphorylation and their role in the mechanisms of formations of functional complexes. The Agard Lab combined cellular cryo-ET and AlphaFold2 modeling to build a workflow to identify proteins [11], track their localization, and determine their structures with the goal of understanding how mammalian sperms are built *in situ*. Notably, their cellular cryo-ET and subtomogram averaging provided 6.0 Å reconstructions of axonemal microtubule structures. Tertiary structures turned out to be well resolved at this resolution allowing the authors to unbiasedly match sperm-specific densities with 21,615 AlphaFold2-predicted protein models of the mouse proteome. They identified novel microtubule-associated proteins forming an extensive interaction network, which led to suggest a role for them in determining the mechanical properties of the filaments.

Despite this sophistication, there is still no experimental technique that can provide insight at an atomic level into the dynamic processes underlying recognition in multiprotein assembly formation nor set the stage for the definition of rules for the design of molecules able to modulate/perturb functional pathways. To understand complex biology at an atomic level, we have little choice but to turn to theoretical/computational approaches.

Here, we briefly review the main advances and applications in the study of the connections among structure-dynamics-recognition and function in protein systems, and discuss how this knowledge can be leveraged to study and predict complex supramolecular structures, and ultimately design assembly-specific interactors. In this framework, we discuss advancements in the use of classical molecular dynamics (MD) simulations and then we extend our attention to advanced methods based on machine and deep learning.

The dynamics underlying protein molecular recognition and function

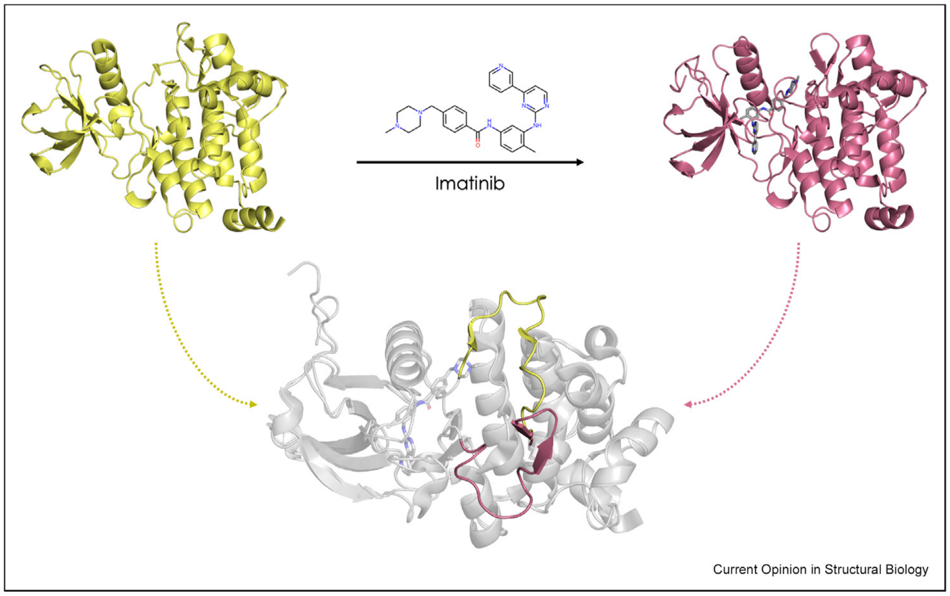
Understanding the mechanisms of protein-drug and protein-protein association is the first fundamental step for the realistic description of biochemical phenomena.

Ayaz et al. [12] used extensive MD simulations to obtain an atomic level description of the protein conformational changes that take place when small molecules bind. Specifically, the authors set out to study imatinib binding to Abl kinase. Through unguided long timescale simulations, they showed the drug first selectively binding the autoinhibitory conformation of the kinase. After this step, imatinib induces a large conformational change of the protein to reach a bound state comparable to the published crystal structures. The reliability of unbiased simulations in returning experimentally consistent aspects of a complex binding pathway is only one aspect of this paper. Indeed, the authors reveal an unexpected local structural instability (cracking) in the C-terminal lobe of Abl kinase, which emerges during the binding process. The region corresponds to a substructure where mutations conferring drug resistance tend to accumulate. The computational predictions are then used to design mimics of resistant mutants. nuclear magnetic resonance (NMR) spectra, hydrogen—deuterium exchange measurements, and thermostability measurements on the mutants suggest that these mutations confer imatinib resistance indeed by exacerbating structural instability in the C-terminal lobe: the end effect of mutations is to render the drug-bound state energetically unfavorable (Figure 1).

Shedding light on the mechanisms of protein-protein complex formation as well as predicting their kinetics is a challenging task. To this end, microscopic modeling of association and dissociation events is a requirement, which has often been hampered by the lack of efficient sampling methods. Noè et al. [13,14] combined high-throughput adaptive molecular dynamics (MD) simulations with Markov modeling to study the association between ribonuclease barnase and its inhibitor barstar. Notably, they showed the possibility to access experimentally consistent intermediate structures, revealing an ensemble of transient (mis-bound) states and a funnel-shaped energy landscape driving the sampled complexes to the native basin. Notably, the use of Markov models allowed to obtain a quantitative profiling of the kinetics on the microsecond to hours timescales (Figure 2).

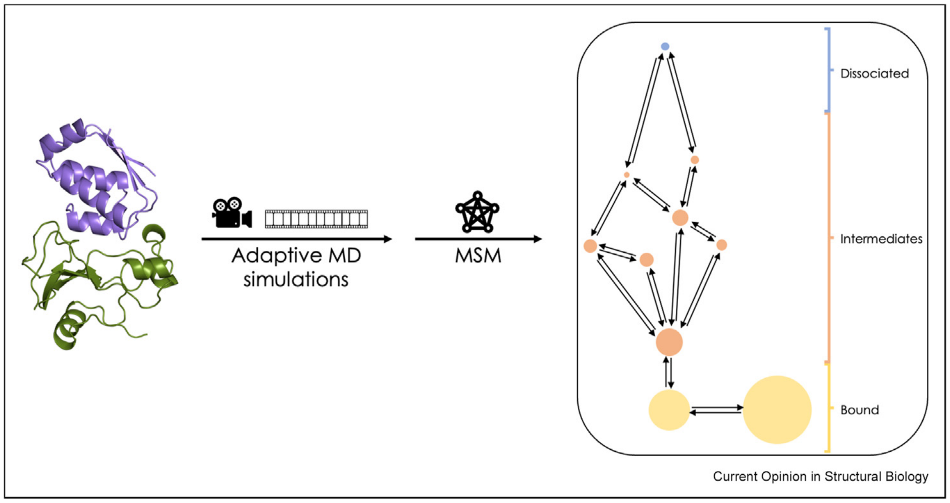
An important contribution to the study of molecular encounters and mechanisms leading to the formation of multiprotein complexes comes from Brownian Dynamics (BD) simulations. Key work in this area is discussed in Ref. [15]. While BD has been used as a

Figure 1



The conformational changes from the apo form of Bcr-Abl kinase (yellow) to the bound state (pink) with the inhibitor Imatinib (above the arrow). In the lower part of the figure is highlighted the movement of the activation loop in the apo and bound state (respectively in yellow and pink).

Figure 2



A representative scheme of the study on the ribonuclease barnase in a multiprotein complex with the barstar inhibitor. The combination of adaptive MD simulation and Markov State Models analysis allows determining intermediate states from the bound complex to the isolated proteins.

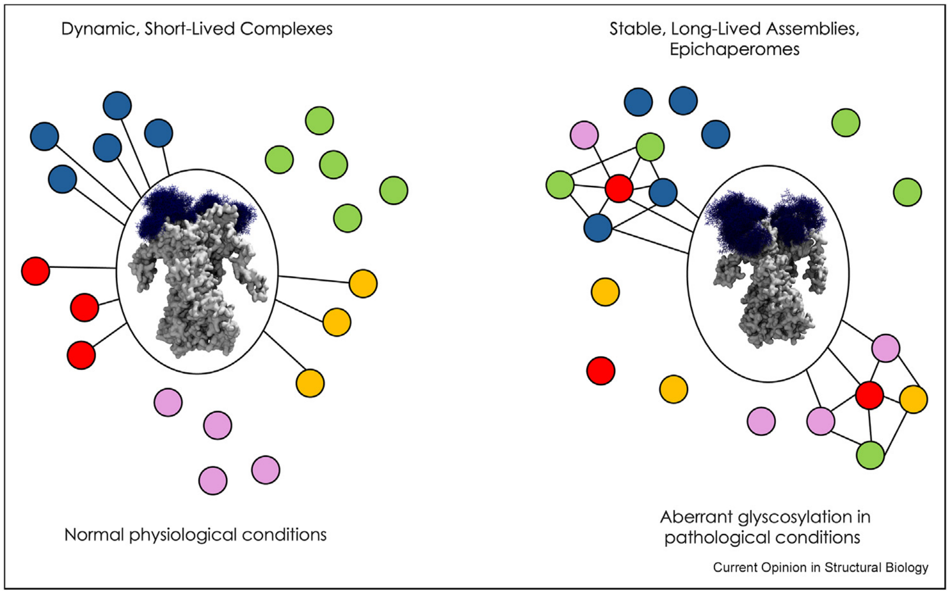
computational tool to investigate molecular diffusion, the latest developments in methodologies and improvements in computing power are extending the reach, scope and application ranges of this simulative approach. Indeed, BD has been used successfully to compute association rate constants and generate possible structures complexes between binding partners (protein-protein or protein-ligand interactions). In BD, a number of simulations is performed, and diffusional and kinetic properties are calculated. In this scenario, biomolecules are commonly simulated at atomistic resolution, although coarse-graining and multiscale representations are increasingly being used. Furthermore, mutants of involved partners can also be considered to evaluate both their effects on the mechanisms of formation and the structures of the resulting complexes.

Extensive simulations have also been recently used to probe the dynamic mechanisms that control the selection of protein conformational states that are then presented for interactions with partners depending on the state of the cell, as well as the mechanisms of allosteric cross-talk between recognition sites [16,17]. In this context, simulative approaches have been used to model the impact of aberrant post-translational modifications on the activity and recognition mechanisms of a central regulator of protein homeostasis, namely chaperone protein GRP94. Specifically, we rationalized the impact of the pathologic N-glycosylation on Asn62 compared with the physiologic one, namely glycosylation at Asn217, to show that each post translational modification (PTM) induces distinct states of GRP94, which are selectively poised/preorganized to interact with distinct pools of proteins (Figure 3) [18,19].

MD-based structural knowledge can further be leveraged to develop models of functional multiprotein assemblies. In this context, Mysore et al. [20] used MD-generated models of K-Ras, a fundamental regulator of MAPK pathways in cell growth implicated in many cancers, to build atomistic models of multiple K-Ras assemblies at the cell membrane, and shed light on its interaction with Ras effector proteins. The starting point was an asymmetric guanosine triphosphate-mediated K-Ras dimer model. Adding further K-Ras monomers to this initial nucleus in a head-to-tail fashion led to a compact helical assembly. Importantly, the model was validated using both electron microscopy and cell-based experiments. Results indicate that K-Ras can be stabilized in its active state, while at the same time presenting the correct interfaces to recruit Raf. Using experimentally based constraints, the authors positioned C-Raf, kinase MEK1, and Galectin-3 and 14-3-3σ on and around the helical assembly. The model is finally shown to provide a structural basis to rationalize a large body of data on MAPK signaling.

Hoff and Bonomi [21] exploited the description of protein conformational heterogeneity obtainable from MD simulations to improve structural reconstruction from Cryo-EM data. Importantly, in some cases the flexibility of interacting regions results in low resolution, in averaging out conformational details, and in an inability to accurately determine local structures. The authors here introduce a Bayesian inference approach to determine structural ensembles of biological entities by combining cryo-EM data with molecular dynamics (MD) simulations. Their approach automatically detects and downweighs noisy experimental data, calculates accurate structural ensembles of proteins and protein complexes including any lipids, small molecules and ordered water present in experimental maps. The method, called EMMIvox is benchmarked against a number of known cases, showing a clear improvement in resolution and then applied to define the structural ensembles of the

Figure 3


Dynamic, Short-Lived Complexes
Stable, Long-Lived Assemblies, Epichaperomes
Normal physiological conditions
Aberrant glyscosylation in pathological conditions
Current Opinion in Structural Biology

A schematic representation of the chaperone heat shock protein Grp94 (in gray surface) at different levels of glycosylation (glycans in blue sticks). The glycan density represents different conformations sampled along the MD trajectory simulations. The two schemes depict the different behaviors of the protein in normal conditions, left panel, in disease conditions, and the right panel, in determining the assembly of short- vs. long- lived assemblies, respectively (see Ref. [19]).

type 1a tau filament (1.9 Å) and the SPP1 bacteriophage (4 Å) in detail. Importantly, EMMIVox is made available through the free PLUMED library.

Synergizing AI and MD to enhance the reach of simulations

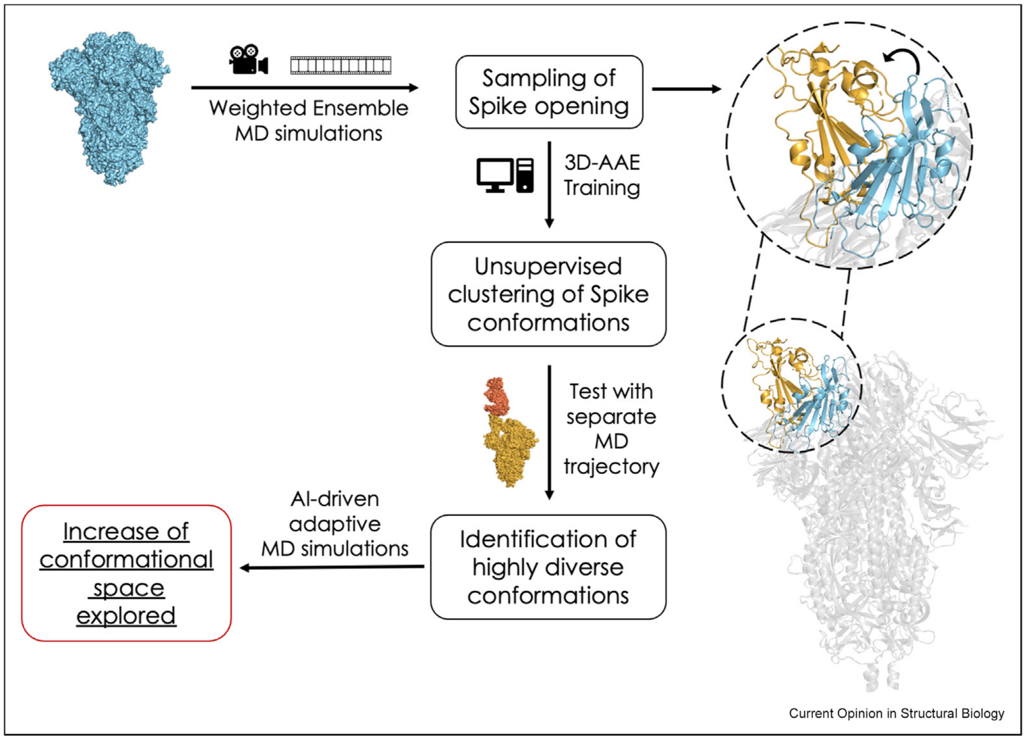
The examples reported above show that it is now possible to garner high resolution insights into complex mechanisms and improve the quality of experimental models of large assemblies using atomistic simulations. However, these studies are still somewhat limited by the necessity to run a large number of simulations and to analyze them with methods suitable for the specific scientific questions at hand.

In this context, the use of AI to learn functional properties of complex systems from simulations in an automated fashion can potentially transform the use and impact of computational biology [22].

Amaro et al. [23,24] demonstrated the feasibility of exploiting generalizable AI-driven workflows to extract information from simulation data that originate from heterogeneous high performance computing (HPC) resources. In their studies, AI-based workflows are used to explore the time-dependent dynamics and the mechanisms of infectivity of the SARS-CoV-2 spike protein, enabling efficient investigation of spike motions in several complex environments. Notably, these include a complete SARS-CoV-2 viral envelope simulation, which entails calculations on a system of 305 million atoms. AI techniques are integrated with the weighted ensemble method [25], a splitting strategy that replicates promising MD trajectories to increase the sampling efficiency of complex and rare events. In this framework, unsupervised linear and non-linear dimensionality reduction identify collective reaction coordinates from high-dimensional systems, which advantageously guide a system (e.g. the fully glycosylated spike protein) through conformational transitions (e.g. from the closed to open states). Interestingly, the authors show the capacity of these AI approaches to automatically classify and stratify reaction coordinates (Figure 4).

Significant progress in using AI in MD simulations comes from initiatives aimed at learning effective interaction potentials via machine or deep learning. This entails developing (deep) neural networks to

Figure 4



A scheme of the combination of weighted ensemble MD simulations with an artificial intelligence-based approach to investigate the flexibility of the SARS-CoV-2 spike protein. In the right image it highlighted the opening of the receptor binding domain (RBD) domain switching from the closed state (light blue) to the open state (yellow).

predict quantum mechanical energies and forces, to generate coarse-grained representations for the simulation of large systems, or to sample equilibrium structures while computing thermodynamic properties at the same time [26]. An interesting example focused on sampling diverse conformations and the dynamics of complex mechanisms in the context of the protein folding problem is represented by the work of Majewski et al. [27]. The authors train neural networks using data from unbiased all-atom MD simulations of twelve distinct systems (about 9 ms of sampling overall): the training set spans differential secondary structure arrangements, a number of folding-unfolding transitions, as well as an extensive sampling of native and misfolded states. The neural network potential learned from this data set shows the ability to accelerate the dynamics by more than three orders of magnitude, while preserving the thermodynamics of the systems. markov state model (MSM) analysis of the coarse-grained trajectories showed that all the protein models were able to recover the respective experimental native structure of the corresponding target, also predicting the evolution of secondary and tertiary structures. Importantly, a single coarse-grained potential can integrate all twelve proteins and recapitulate experimental structural features of mutated proteins.

Tiwary et al. [28] combined AlphaFold2 and MD simulations potentiated with AI-driven enhanced sampling to generate Boltzmann-weighted conformational ensembles from sequence. The method is called AlphaFold2-RAVE. Their protocol first uses a reduced multiple sequence alignment to induce AlphaFold2 to generate many possible conformations as the starting structure for subsequent enhanced sampling. The method is based on learning appropriate reaction coordinates for correctly sampling conformational states. The authors demonstrate that the method returns a Boltzmann-weighted ensemble of protein conformations. As a key applicative example, AlphaFold2-RAVE [29] is applied to characterize the conformational landscape, the Asp-Phe-Gly (DFG) loop of kinase (DDR1). In this framework, we note that AlphaFold2 can proficiently be used to sample different, physically meaningful conformations using a subsampling of the initial MSA [30]. This subsampling approach is shown to be particularly useful to predict the impact of mutations on the conformational landscape and well-populated states of proteins.

Wellawatte et al. [31] compared neural network (NN)-based coarse-grained force fields to traditional coarse-grained force fields. The authors interestingly show that NN force fields are able to extrapolate to unseen regions of the free energy surface upon training with limited data sets, supporting the generalized applicability of these kinds of methods.

While promising, the possibility for AI-based approaches to recover conformational ensembles with the correct Boltzmann weights still present a number of limitations and critical points. Some of these may also reflect the known limitations of underlying MD simulations in efficiently sampling complex landscapes. For an in-depth discussion of these issues, we refer the readers to this interesting review by Mchaourab et al. [32].

When simulating high-dimensional phenomena, identifying a reduced set of collective variables that recapitulate their key physical determinants can be critical for two reasons: on the one hand, it may help gain a better understanding of atomistic simulations, while on the other hand it may favor accelerating simulations through integration with enhanced sampling techniques. The Parrinello et al. developed a series of theoretical and computational tools to learn these variables directly from atomistic data [33—35]. The learning processes entail deep targeted discriminant analysis of data from short unbiased simulations and transition path ensembles, dimensionality reduction and classification of metastable states [35], or identification of slow modes [33].

Finally, we note here that the combination of MD methods can be combined with AI to streamline drug design and make it more efficient, in particular, the screening of large small-molecule databases [36]. In this context, free energy calculations can be used in the so-called active-learning free energy perturbation approaches [37]. Here, predictions on one central reference compound from a random subset are used to train a machine learning (ML) model. The resulting model is used to score the remaining compounds in the library. From the obtained ML scores, the next set of compounds is selected for a round of FEP profiling, and an improved ML model is trained based on the cumulative predictions. This cycle is then iteratively implemented until all promising compounds have been retrieved from the library. These approaches are actually being showcased for the rapid exploration of large chemical spaces aimed, and in lead optimization [38].

As an example of integration of MD and ML, Goβen et al. identified ligand candidates with novel chemotypes while preserving antagonistic potential and affinity in the nanomolar to target G protein-coupled receptors (GPCRs) [39]. Their approach integrates structural data with a random forest agonist/antagonist classifier and a signal-transduction kinetic model.

In the context of drug discovery, MD/AI approaches can be used to identify cryptic pockets, not immediately evident from the static 3D structures of target proteins [40].

Learning the structures of assemblies

The progress described in the previous paragraphs provides a glimpse into the opportunities generated by the integration of multiple computational methods for understanding complex mechanisms in biology. The advent of AlphaFold2 [41] (AF2) has clearly revolutionized the field of structural studies. Further progress of AF2 into AlphaFold Multimer [42] and AlphaMissense [43] is now allowing pushing the structural detail into the description of complexes involved in signaling pathways (Figure 5).

The discrimination of correct (or functional) models in predicted complexes, a large number of which can now be generated by ML/DL combined with Docking methods, is a key issue for the actual usability of computational predictions. In this context, DeepRank-GNN [44] is an interesting application that, starting from the conversion of 3D structural information on protein-protein interfaces into graphs, is able to learn specific interaction patterns. DeepRank-GNN proves efficient in scoring docking poses and in discriminating biological and crystal interfaces.

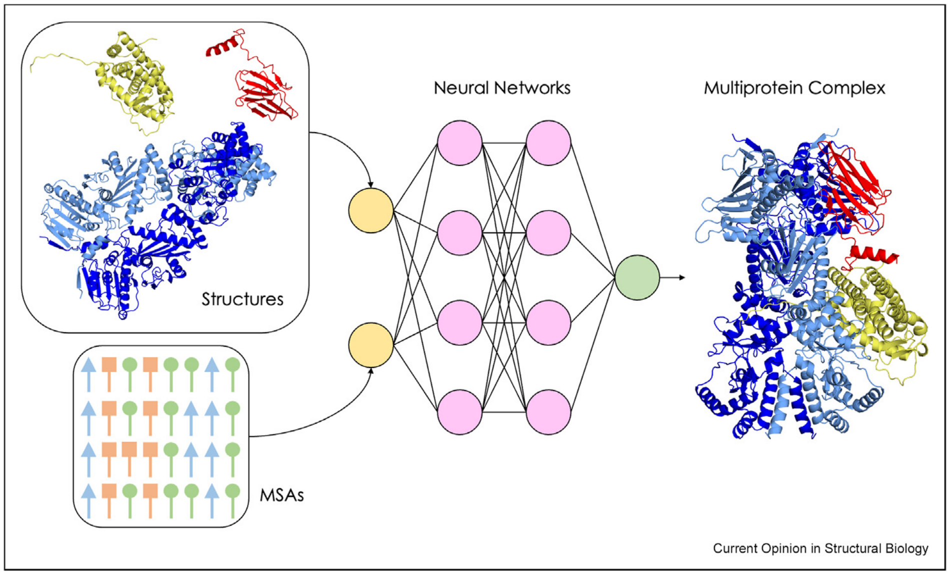
Bryant et al. [45] applied the AF2 protocol with an optimized multiple sequence alignment (MSA) strategy to generate high quality models of heterodimers. From the predicted interfaces, the authors evolve a simple function to predict the acceptable vs. incorrect models as well as interacting with non-interacting proteins.

Skolnick [46] et al. introduced a significant improvement over AF-Multimer, called AF2Complex. This uses the same neural network models as AF2 in single-chain prediction, adapted for multimeric complexes. Significantly, the authors show the possibility to do this without retraining. The authors also devise metrics to predict the probabilities of protein-protein interactions across diverse protein pairs. Following through validation on the *E. Coli* proteome, they were able to build up high confidence models of three complexes, made up of 8 partners, involved in system I for electron transfer and respiration.

Finally, Jandova et al. [47] demonstrated the validity of a simple protocol based on short MD simulations combined with ML (the random forest classifier) to correctly assign native complexes from a number of HADDOCK-generated solutions. Interestingly, native models showed higher stability in almost every measured property, including the key ones used for scoring in the Critical Assessment of Predicted Interaction competition.

Bryant and Noe [48] recently introduced a multistep method that allows decreasing the computational time and storage needs required to build a full PPI network. First, they use a reduced MSA creation procedure and then they pair MSA using species information preserving the single chain data through block diagonalization. Afterward, AlphaFold2 is used for structure prediction, and the predictions undergo evaluation using a scoring

Figure 5



A simplified representation of the use of AlphaFold2 Multimer to predict the structure of a multiprotein complex, starting from multiple sequence alignments (MSAs) to structure of the single components.

scheme. To speed up the processing, they employed a combination of CPU and GPUs, making this procedure parallelizable, as well as freely available.

Overall, these examples demonstrate the possibility of expanding the use of AF2 well beyond the domains and datasets on which it has been based and trained.

Perspectives and conclusions

Recent technological advances are flooding the community with an unprecedented amount of data on protein structures, dynamics, interactions, and functions at different levels of resolution. Accessible data range from the 3D structures of single proteins and protein complexes, to binding affinity and kinetic characterizations, to the definition of signaling pathways and interacting networks at the whole proteome level and in different cellular states. This wealth of data is making it possible to use learning (machine or deep learning) algorithms to extract useful information on complex systems, while developing approaches that can generalize tasks on which they had not specifically been trained on. One key development in this realm is the possibility to learn (or automatically develop) molecular design rules for the generation of proteins with desired functions.

Examples of such endeavors are starting to appear in literature: Language models trained on the protein sequence space have shown the ability to generate *de novo* protein sequences that are distantly related to natural ones and whose properties mirror them in terms of globularity and amount of disorder [49], when characterized experimentally with structural analyses. Wang et al. [50] describe two deep-learning methods to design proteins and enzymes by scaffolding specific binding/active sites required for a certain function using a 3D-structure obtained without the need to prespecify a certain fold or secondary structure. In the first approach, they identify sequences predicted to fold into 3D structures that host the functional site. In the second approach, which they call "inpainting," the design starts from the functional site and adds in the sequence and structure of a prospectively viable protein scaffold using a retrained RoseTTAFold network. The variety of the designs and functions strongly supports the applicability of these methods to develop biomolecules with new (non-natural) functions as biological drugs, novel catalysts, scaffolds. Finally, Watson et al. [51] developed a generative model of protein backbones to design protein monomers, binders, and symmetric oligomers. This method, called RoseTTAFold diffusion (RFdiffusion), is validated experimentally on a number of designed proteins, whose actual 3D structures are demonstrated to be highly similar to the computed ones. Results indicate that state-of-the-art learning approaches have reached levels of success that were to some degree unexpected or at least not fully predictable.

However, ML methods can still be ameliorated and improved to generate predictions outside their domain of input data. In this context, physics- and chemistry-based simulative approaches can be integrated into the training and development processes to generate realistic information on interaction propensities, structural stabilities, to predict affinities, or to access parts of the conformational space of a protein that can be important for function, but are not immediately evident from available structural data.

It is important to point out here that there is still plenty of room for directly using physical chemistry-based simulation methods to investigate biological structures and functions, even those generated by ML. Indeed, MD can provide direct access to information on functionally-oriented dynamic properties, on the structural characteristics of transition state ensembles, and even on the relevance of specific post-translational modifications.

Furthermore, many phenomena in biology require the use of quantum mechanics or mixed quantum mechanics/molecular mechanics approaches to access realistic mechanistic information [52]. These phenomena include, among others, enzyme catalysis, electron-transfer mechanisms, recognition, and transport phenomena where charge transfer or charge correlations are key [53,54], in particular where metal ions are involved. The advent of the massively parallel supercomputers is ushering computational biology and chemistry into the exascale era. This new dimension and the opportunities it generates for massive calculations are promising to give an incredible boost to the use of quantum-based techniques for instance in the field of drug design [55]. On the one hand, these additional capabilities will permit to accurately model highly complex systems at multiple and larger scales [14]. On the other hand, they will expectedly generate massive amounts of data of high quality useful for the training of novel ML/AI algorithms to find patterns in complex reactive mechanisms, predict potential reaction pathways, intermediates, transition states, and sample conformational and other transitions between distinct states in large and biologically realistic assemblies [54].

All of these cases exemplify that while in some instances simulations can unveil knowledge that is still out of reach for most ML/DL-based strategies, the possibilities to merge these worlds are starting to become more and more evident.

In this frame of thought, it is evident that there exists a wide space for exploring and training new methods on a broader spectrum of functional propensities, PPIs, and information on biologically relevant assemblies, which remain key features to understand and rationalize the exact functioning of complex molecular organizations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

The authors acknowledge funding from Fondazione AIRC (Associazione Italiana Ricerca Sul Cancro) under IG 2022 - ID. 27139; from PRIN (grant 20209KYCH9); from IMMUNO-HUB, T4-CN-02, from the Ministero della Salute (Italy); Programma di ricerca PNRR CN00000013 “National Centre for HPC, Big Data and Quantum Computing”.

References

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Wei G, Xi W, Nussinov R, Ma B: **Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell**. *Chem Rev* 2016, **116**:6516–6551.

2. • Nussinov R, Tsai C-J, Jang H: **Protein ensembles link genotype to phenotype**. *PLoS Comput Biol* 2019, **15**. e1006648-e1006648.

This paper presents an important framework whereby protein structures are viewed as the connection between genotypes and phenotypes in cellular pathways.

3. Nussinov R, Tsai C-J, Jang H: **Signaling in the crowded cell**. *Curr Opin Struct Biol* 2021, **71**:43–50.

4. Paananen J, Fortino V: **An omics perspective on drug target discovery platforms**. *Briefings Bioinf* 2020, **21**:1937–1953.

5. Pricer R, Gestwicki JE, Mapp AK: **From fuzzy to function: the new frontier of protein-protein interactions**. *Accounts Chem Res* 2017, **50**:584–589.

6. Gestwicki JE, Shao H: **Inhibitors and chemical probes for molecular chaperone networks**. *J Biol Chem* 2019, **294**: 2151–2161.

7. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A: **Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies**. *PLoS Biol* 2012, **10**, e1001244.

8. Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull M-T, *et al.*: **AI-based structure prediction empowers integrative structural analysis of human nuclear pores**. *Science* 2022, **376**, eabm9506.

9. Digvijay S, Neelesh S, Joshua H, Ignacia E, Farhaz S, Madeleine D, Sergey S, Zhixun L, Trevor van E, Kelly M, *et al.*: **The molecular architecture of the nuclear basket**. *bioRxiv* 2024. 2024.2003.2027.587068.

10. Juyoux P, Galdadas I, Gobbo D, von Velsen J, Pelosse M, Tully M, Vadas O, Gervasio FL, Pellegrini E, Bowler MW: **Architecture of the MKK6-p38α complex defines the basis of MAPK specificity and activation**. *Science* 2023, **381**: 1217–1225.

11. Chen Z, Shiozaki M, Haas KM, Skinner WM, Zhao S, Guo C, Polacco BJ, Yu Z, Krogan NJ, Lishko PV, *et al.*: **De novo protein identification in mammalian sperm using in situ cryoelectron tomography and AlphaFold2 docking**. *Cell* 2023, **186**: 5041–5053.e5019.

12. • Ayaz P, Lyczek A, Paung Y, Mingione VR, Iacob RE, de Waal PW, Engen JR, Seeliger MA, Shan Y, Shaw DE: **Structural mechanism of a drug-binding process involving a large conformational change of the protein target**. *Nat Commun* 2023, **14**:1885.

This paper shows at atomistic level the mechanisms of binding and the sequence of conformational transition events linked to the recognition of the binding-site by the ligand.

13. Plattner N, Noé F: **Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models**. *Nat Commun* 2015, **6**:7653. 7653.

14. •• Plattner N, Doerr S, De Fabritiis G, Noé F: **Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling**. *Nat Chem* 2017, **9**:1005–1011.

This paper determines the full kinetics and mechanisms of protein-protein complex formation, using fully atomistic models.

15. Muñiz-Chicharro A, Votapka LW, Amaro RE, Wade RC: **Brownian dynamics simulations of biomolecular diffusional association processes**. *WIREs Computational Molecular Science* 2023, **13**, e1649.

16. Aguti R, Bernetti M, Bosio S, Decherchi S, Cavalli A: **On the allosteric puzzle and pocket crosstalk through computational means**. *J Chem Phys* 2023, **158**, 165101.

17. Castelli M, Magni A, Bonollo G, Pavoni S, Frigerio F, Oliveira ASF, Cinquini F, Serapian SA, Colombo G: **Molecular mechanisms of chaperone-directed protein folding: insights from atomistic simulations**. *Protein Sci* 2024, **33**, e4880.

18. •• Castelli M, Yan P, Rodina A, Digwal CS, Panchal P, Chiosis G, Moroni E, Colombo G: **How aberrant N-glycosylation can alter protein functionality and ligand binding: an atomistic view**. *Structure* 2023, **31**:987–1004.e1008.

The authors show how Post-Translational Modifications impact on the conformational profile of a protein specifically in disease states. The information obtained is exploited to explain the function and activity of active vs. inactive drugs.

19. Chiosis G, Digwal CS, Trepel JB, Neckers L: **Structural and functional complexity of HSP90 in cellular homeostasis and disease**. *Nat Rev Mol Cell Biol* 2023, **24**:797–815.

20. Mysore VP, Zhou Z-W, Ambrogio C, Li L, Kapp JN, Lu C, Wang Q, Tucker MR, Okoro JJ, Nagy-Davidescu G, *et al.*: **A structural model of a Ras–Raf signalosome**. *Nat Struct Mol Biol* 2021, **28**: 847–857.

21. Hoff S, Bonomi M: **A Bayesian inference approach to determining structural ensembles using cryo-EM and molecular dynamics**. *Biophys J* 2023, **122**, 180a.

22. Glielmo A, Husic BE, Rodriguez A, Clementi C, Noé F, Laio A: **Unsupervised learning methods for molecular simulation data**. *Chem Rev* 2021, **121**:9722–9758.

23. Casalino L, Dommer AC, Gaieb Z, Barros EP, Sztain T, Ahn S-H, Trifan A, Brace A, Bogetti AT, Clyde A, *et al.*: **AI-driven multi-scale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics**. *Int J High Perform Comput Appl* 2021, **35**: 432–451.

24. Dommer A, Casalino L, Kearns F, Rosenfeld M, Wauer N, Ahn S-H, Russo J, Oliveira S, Morris C, Bogetti A, *et al.*: **#COVIDisAirborne: AI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol**. *Int J High Perform Comput Appl* 2022, **37**:28–44.

25. Huber GA, Kim S: **Weighted-ensemble Brownian dynamics simulations for protein association reactions**. *Biophys J* 1996, **70**:97–110.

26. Noé F, Tkatchenko A, Müller K-R, Clementi C: **Machine learning for molecular simulation**. *Annu Rev Phys Chem* 2020, **71**: 361–390.

27. • Majewski M, Pérez A, Thölke P, Doerr S, Charron NE, Giorgino T, Husic BE, Clementi C, Noé F, De Fabritiis G: **Machine learning coarse-grained potentials of protein thermodynamics**. *Nat Commun* 2023, **14**:5739.

This paper reports on the development of a general coarse-grained force field able to reproduce the folding thermodynamics of small proteins.

28. Vani BP, Aranganathan A, Wang D, Tiwary P: **AlphaFold2-RAVE: from sequence to Boltzmann ranking**. *J Chem Theor Comput* 2023, **19**:4351–4354.

29. Vani BP, Aranganathan A, Tiwary P: **Exploring kinase asp-Phe-Gly (DFG) loop conformational stability with AlphaFold2-RAVE**. *J Chem Inf Model* 2024, **64**: 2789–2797.

30. Monteiro da Silva G, Cui JY, Dalgarno DC, Lisi GP, Rubenstein BM: **High-throughput prediction of protein conformational distributions with subsampled AlphaFold2**. *Nat Commun* 2024, **15**:2464.

31. Wellawatte GP, Hocky GM, White AD: **Neural potentials of proteins extrapolate beyond training data**. *J Chem Phys* 2023, **159**, 085103.

32. Brown BP, Stein RA, Meiler J, McHaourab HS: **Approximating projections of conformational Boltzmann distributions with AlphaFold2 predictions: opportunities and limitations**. *J Chem Theor Comput* 2024, **20**: 1434–1447.

33. Novelli P, Bonati L, Pontil M, Parrinello M: **Characterizing metastable states with the help of machine learning**. *J Chem Theor Comput* 2022, **18**:5195–5202.

34. Bonati L, Trizio E, Rizzi A, Parrinello M: **A unified framework for machine learning collective variables for enhanced sampling simulations: mlcolvar**. *J Chem Phys* 2023, **159**, 014801.

35. Ray D, Trizio E, Parrinello M: **Deep learning collective variables from transition path ensemble**. *J Chem Phys* 2023, **158**, 204102.

36. Sadybekov AV, Katritch V: **Computational approaches streamlining drug discovery**. *Nature* 2023, **616**:673–685.

37. Knight JL, Leswing K, Bos PH, Wang L: **Impacting drug discovery projects with large-scale enumerations, machine learning strategies, and free-energy predictions**. In *Free energy methods in drug discovery: current state and future directions*. Edited by Chemical Society American, *ACS symposium series*, **1397**; 2021:205–226.

38. Bos PH, Houang EM, Ranalli F, Leffler AE, Boyles NA, Eyrich VA, Luria Y, Katz D, Tang H, Abel R, *et al.*: **AutoDesigner, a de novo design algorithm for rapidly exploring large chemical space for lead optimization: application to the design and synthesis of d-amino acid oxidase inhibitors**. *J Chem Inf Model* 2022, **62**: 1905–1915.

39. Goßen J, Ribeiro RP, Bier D, Neumaier B, Carloni P, Giorgetti A, Rossetti G: **AI-based identification of therapeutic agents targeting GPCRs: introducing ligand type classifiers and systems biology**. *Chem Sci* 2023, **14**:8651–8661.

40. Meller A, Bhakat S, Solieva S, Bowman GR: **Accelerating cryptic pocket discovery using AlphaFold**. *J Chem Theor Comput* 2023.

41. •• Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al.*: **Highly accurate protein structure prediction for the human proteome**. *Nature* 2021, **596**:590–596.

The paper describes AlphaFold2, which has revolutionized the field of protein structure prediction and opened the door to its exploitation in the generation of biologically valuable complexes.

42. Richard E, Michael ON, Alexander P, Natasha A, Andrew S, Tim G, Augustin Ž, Russ B, Sam B, Jason Y, *et al.*: **Protein complex prediction with AlphaFold-Multimer**. *bioRxiv* 2022. 2021.2010.2004.463034.

43. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, *et al.*: **Accurate proteome-wide missense variant effect prediction with AlphaMissense**. *Science* 2023, **381**, eadg7492.

44. Réau M, Renaud N, Xue LC, Bonvin AMJJ: **DeepRank-GNN: a graph neural network framework to learn patterns in protein–protein interfaces**. *Bioinformatics* 2023, **39**, btac759.

45. • Bryant P, Pozzati G, Elofsson A: **Improved prediction of protein-protein interactions using AlphaFold2**. *Nat Commun* 2022, **13**:1265.

Here, the new potential of AlphaFold2 is exploited to generate protein-protein complexes with high accuracy.

46. Gao M, Nakajima An D, Parks JM, Skolnick J: **AF2Complex predicts direct physical interactions in multimeric proteins with deep learning**. *Nat Commun* 2022, **13**:1744.

47. Jandova Z, Vargiu AV, Bonvin AMJJ: **Native or non-native protein–protein docking models? Molecular dynamics to the rescue**. *J Chem Theor Comput* 2021, **17**:5944–5954.

48. Patrick B, Frank N: **Rapid protein-protein interaction network creation from multiple sequence alignments with Deep Learning**. *bioRxiv* 2023. 2023.2004.2015.536993.

49. Ferruz N, Schmidt S, Höcker B: **ProtGPT2 is a deep unsupervised language model for protein design**. *Nat Commun* 2022, **13**:4348.

50. •• Wang J, Lisanza S, Juergens D, Tischer D, Watson JL, Castro KM, Ragotte R, Saragovi A, Milles LF, Baek M, *et al.*: **Scaffolding protein functional sites using deep learning**. *Science* 2022, **377**:387–394.

The authors describe two deep-learning methods to design proteins with tailored functional sites. The authors prove the validity of their approaches via the design of proteins containing a variety of functional motifs.

51. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, Ahern W, Borst AJ, Ragotte RJ, Milles LF, *et al.*: **De novo design of protein structure and function with RFdiffusion**. *Nature* 2023, **620**:1089–1100.

52. Serapian SA, Moroni E, Ferraro M, Colombo G: **Atomistic simulations of the mechanisms of the poorly catalytic mitochondrial chaperone Trap1: insights into the effects of structural asymmetry on reactivity**. *ACS Catal* 2021, **11**: 8605–8620.

53. Olsen JMH, Bolnykh V, Meloni S, Ippoliti E, Bircher MP, Carloni P, Rothlisberger U: **MiMiC: a novel framework for multiscale modeling in computational chemistry**. *J Chem Theor Comput* 2019, **15**:3810–3823.

54. Manathunga M, Aktulga HM, Götz AW, Merz Jr KM: **Quantum mechanics/molecular Mechanics simulations on NVIDIA and AMD graphics processing units**. *J Chem Inf Model* 2023, **63**: 711–717.

55. Raghavan B, Paulikat M, Ahmad K, Callea L, Rizzi A, Ippoliti E, Mandelli D, Bonati L, De Vivo M, Carloni P: **Drug design in the exascale era: a perspective from massively parallel QM/MM simulations**. *J Chem Inf Model* 2023, **63**:3647–3658.