Physics-Informed Deep Learning: A New Paradigm for Macromolecular Recognition and Rational Protein Design

Journal:	*The Innovation Drug Discovery*
Manuscript ID	TIDD-2026-0138.R2
Manuscript Type:	Review
Keywords:	Physics-informed deep learning, biomolecular recognition, rational drug discovery, geometric deep learning, protein design
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Main text.tex	

Physics-informed deep learning: A new paradigm for macromolecular recognition and rational protein design

Haobo Xie,[1] Hao Wang,[2] Xiaojun Yao,[3] Jinyan Li,[4] Yi Pan,[4] Aiping Lu,[5] Xiangxiang Zeng,[1] Dongsheng Cao,[6,*] and Xiangzheng Fu[4,*]

[1]College of Computer Science and Electronic Engineering, Hunan University, No. 252 Lushan South Road, Yuelu District, Changsha 410082, Hunan, China
[2]Tobacco Research Institute, Chinese Academy of Agricultural Sciences, Qingdao 266101, China
[3]Faculty of Applied Sciences, Macao Polytechnic University, Macao 999078, China
[4]Computer Science and Artificial Intelligence, Shenzhen University of Advanced Technology, Shenzhen 518107, China
[5]School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR 999077, China
[6]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410003, Hunan, China
[*]Correspondence: oriental-cds@163.com (D.C.); fuxiangzheng@suat-sz.edu.cn (X.F.)

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the [National Natural Science Foundation of China] (Grant Nos. [123456]). **The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.**

AUTHOR CONTRIBUTIONS

H.X., D.C., and X.F. conceived the study and wrote the manuscript. **All authors contributed to the manuscript and approved the final version.**

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

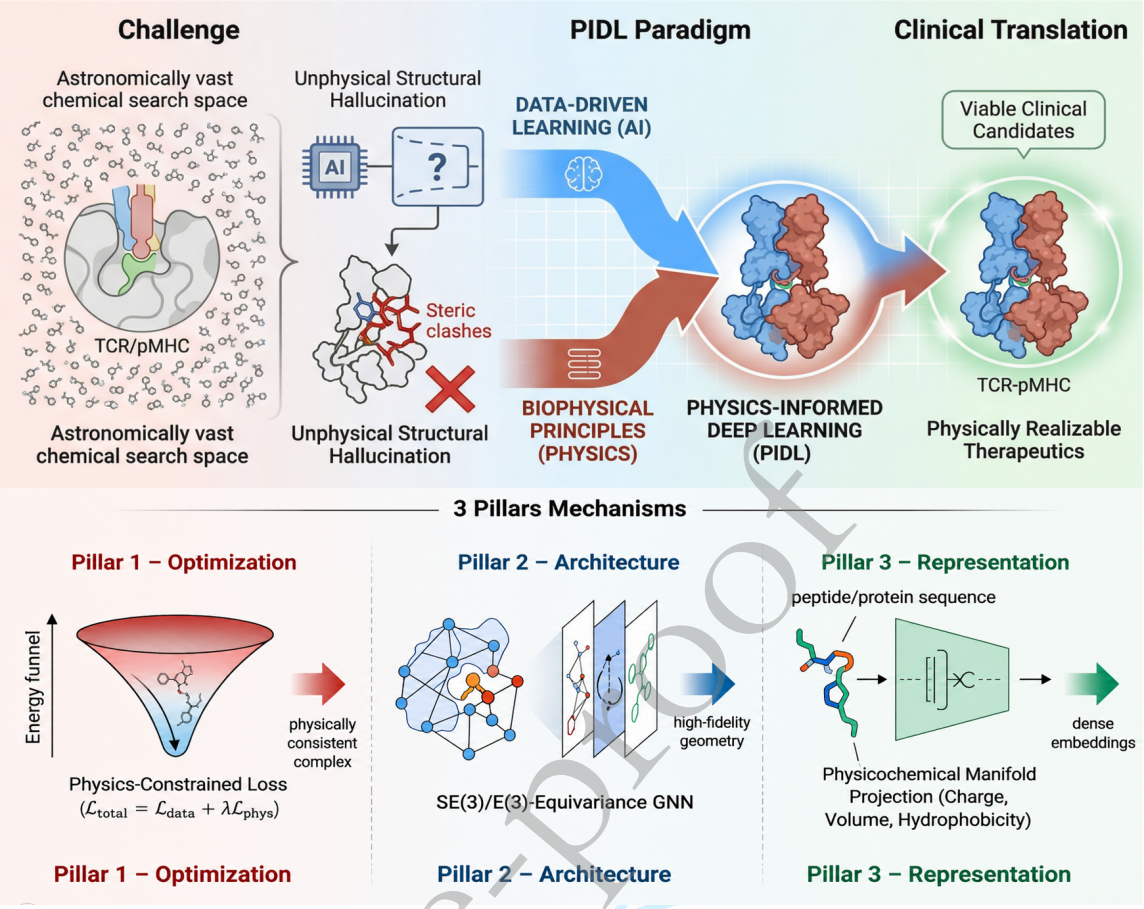
DATA AND CODE AVAILABILITY

No new data were created or analyzed in this study.

ETHICAL STATEMENT

Not Applicable.

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- PIDL grounds biomolecular AI in physical laws to reduce implausible structural predictions.
- Three pillars connect physical constraints, equivariant architectures, and guided representations.
- Physics-aware models improve recognition across ligands, peptides, proteins, and immune complexes.
- Future systems should unite dynamics, cellular context, and closed-loop experimental validation.

ABSTRACT

Targeted drug discovery is fundamentally bottlenecked by the challenge of accurately modeling complex biomolecular interactions, ranging from small-molecule ligand binding to high-order macromolecular assemblies. While traditional physics-based computational methods provide profound mechanistic in-

sights, their clinical utility is frequently hampered by prohibitive computational costs and scalability limitations when addressing highly flexible, cross-scale systems. Conversely, the rapid emergence of pure deep learning offers unprecedented computational speed but suffers from a fundamental "black-box" nature, sometimes yielding physically improbable conformations—often referred to as "hallucinations"—that can pose challenges in real-world experimental validation. To bridge this critical translational gap, the integration of physical principles with artificial intelligence—Physics-Informed Deep Learning (PIDL)—is currently driving a fundamental transition from purely empirical approximations to rational, physically grounded design. This review constructs a strategic framework to critically evaluate these transformative advances, structured around three methodological pillars: (1) Physics-constrained optimization, which integrates thermodynamic principles and integrative experimental restraints at the output level to decode macromolecular dynamics; (2) Physics-encoded architectures, which embed appropriate $SE(3)$ or $E(3)$ geometric symmetries directly into neural network topologies for precise structural recognition; and (3) Physics-guided representations, which project discrete sequences into continuous physicochemical manifolds to enhance interaction prediction. By delineating how physics-based priors synergize with data-driven representation learning, this review not only synthesizes current algorithmic breakthroughs but also provides a comprehensive roadmap for generating physically plausible and thermodynamically stable therapeutics, ultimately accelerating the transition of computationally designed molecules from in silico blueprints to viable clinical candidates.

KEYWORDS: Physics-informed deep learning; biomolecular recognition; rational drug discovery; geometric deep learning; protein design

SUBJECT AREA: Life Sciences

INTRODUCTION

Modern rational drug discovery is a formidable systems engineering challenge, fundamentally bottlenecked by the necessity to accurately model complex biomolecular interactions across an astronomically vast chemical and structural space.[?,?,?,?] Whether predicting the delicate thermodynamic balance of small-molecule ligands binding to dynamic protein pockets, or deciphering the multi-subunit synergy of immune recognition systems like TCR-pMHC complexes, precision at the atomic level remains elusive.[?,?,?,?] Particularly in the realm of tumor immunology, the theoretical diversity of the TCR repertoire exceeds 10^{15}. This immense structural dark matter renders traditional trial-and-error screening both computationally intractable and experimentally cost-prohibitive.[?,?,?]

While highly flexible peptide-protein complexes and immune recognition networks represent the current zenith of modeling difficulty and are heavily featured in this review, the Physics-Informed Deep Learning (PIDL) paradigm proposed herein is fundamentally molecule-agnostic. This methodology seamlessly extends to protein-ligand interactions essential for structure-based drug design and protein-nucleic acid regulatory networks, underscoring the transformative power of embedding universal physical laws over relying on target-specific sequence memorization.

Since the 1970s, Computer-Aided Drug Design (CADD) has evolved from basic Quantitative Structure-Activity Relationship (QSAR) models into sophisticated physics-based structural simulations.[????] Methods such as Molecular Dynamics (MD) and high-resolution molecular docking have enabled the atomic-level elucidation of binding mechanisms, serving as the core engine of traditional rational design.[????] However, in the context of high-throughput screening and ultra-fast generative design, the immense computational cost of methods like Molecular Dynamics (MD), Free Energy Perturbation (FEP), and Quantum Mechanics/Molecular Mechanics (QM/MM) presents significant scalability challenges.[??] Crucially, this does not imply that traditional physics simulations are obsolete. On the contrary, they remain the indispensable gold standard for rigorously validating AI-generated structures, calibrating deep learning energy functions, and elucidating detailed molecular mechanisms at the atomistic or quantum level. The true paradigm shift lies not in replacing physics-based simulations with pure deep learning, but in establishing a highly complementary, symbiotic pipeline. In this modern framework, PIDL serves as the "macroscopic navigator" that rapidly explores the astronomically vast chemical space to generate physically plausible candidate pools, while traditional physics simulations act as the "microscopic verifier" to provide the high-fidelity thermodynamic refinement, kinetic profiling, and mechanistic validation strictly required for clinical progression.

The advent of machine learning, particularly deep learning, has catalyzed a profound paradigm shift, offering data-centric pathways to bypass these computational bottlenecks. Yet, this speed comes at a cost. The black-box nature of pure deep learning frameworks, coupled with their complete thermodynamic blindness, has sparked significant new challenges.[???] Because standard neural networks blindly interpolate statistical correlations without perceiving the underlying physical energy landscapes, they frequently produce structural hallucinationsgeometrically plausible but thermodynamically highly unfavorable conformations often associated with severe steric clashes or solvent-exposed hydrophobic cores.[??]

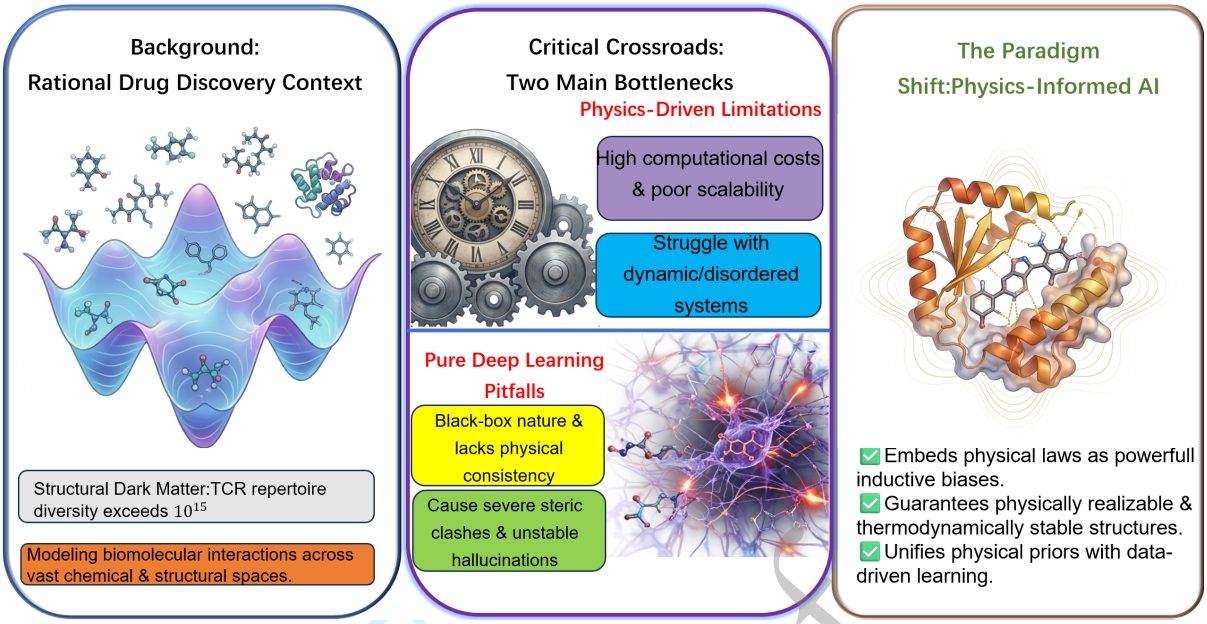


Figure 1. Conceptual prelude: The critical crossroads and paradigm shift in biomolecular modeling The left panel shows the astronomically vast chemical and structural search space in rational drug discovery. The middle panel highlights current bottlenecks, including the high computational costs of traditional physics-driven methods and the unphysical structural hallucinations caused by pure deep learning. The right panel presents Physics-Informed AI as the unifying paradigm, which embeds physical laws to ensure thermodynamically stable and physically realizable predictions.

As conceptualized in Figure 1, the field stands at a critical crossroads. Purely physics-driven models are paralyzed by computational complexity, while purely data-driven models are handicapped by opaque, unphysical predictions. A highly promising strategy lies in their synergistic intersection: Physics-Informed Deep Learning (PIDL).

PIDL represents a theoretical departure from pure empiricism. Rather than forcing a neural network to rediscover the laws of thermodynamics or three-dimensional Euclidean geometry from massive datasets, PIDL mathematically hard-codes established biophysical principles into the algorithmic pipeline as powerful inductive biases. This deep integration effectively constrains the hypothesis space, guiding computational predictions toward physically realizable and clinically viable states. To ensure conceptual rigor, it is imperative to first establish a precise definition of Physics-Informed Deep Learning (PIDL) within the context of biomolecular modeling. We formally define PIDL as a hybrid algorithmic paradigm that systematically integrates established biophysical laws, thermodynamic principles, and geometric symmetries into the data-driven learning process, thereby actively constraining the neural network's hypothesis space. Furthermore, to clarify the conceptual boundaries often blurred in current literature, we explicitly map these varied physical integration strategies into our three foundational pillars based on how they restrict the algorithmic space: (1) **Physics-informed / Physics-inspired Constrained** methods

introduce constraints at the output or optimization level, either through explicit thermodynamic restraints or by mimicking physical phenomena like diffusion; (2) **Physics-encoded** methods impose "hard" geometric constraints mathematically baked directly into the neural network topology (e.g., $SE(3)/E(3)$ Equivariance); and (3) **Physics-guided / Physics-compatible** methods operate at the representation level, explicitly projecting sequences into physical manifolds or utilizing large language models to learn latent spaces that are implicitly compatible with biophysical rules.

To systematically elucidate this transition and avoid terminological ambiguity, this review proposes a comprehensive framework that maps these distinct definitions into three methodological pillars, categorizing them by exactly where physical principles are injected into the modeling pipeline. We first examine physics-constrained optimization in Section 2, exploring how physics-informed "soft" constraints (thermodynamic laws and experimental restraints) reshape the learning objective at the output level to ensure convergence on true free-energy minima. Moving to the internal network topology, Section 3 discusses physics-encoded architectures, highlighting how the "hard" enforcement of spatial symmetries facilitates high-fidelity structural generation. Section 4 traces the evolution of physics-guided representations, detailing the injection of physical priors at the input level from explicit physicochemical manifold projections to implicit biophysical features. Finally, Sections 5 and 6 provide a critical assessment of persistent bottlenecks, such as crystallographic bias and entropic blindness, and outline a strategic roadmap toward context-aware mesoscale modeling and the realization of autonomous discovery laboratories.

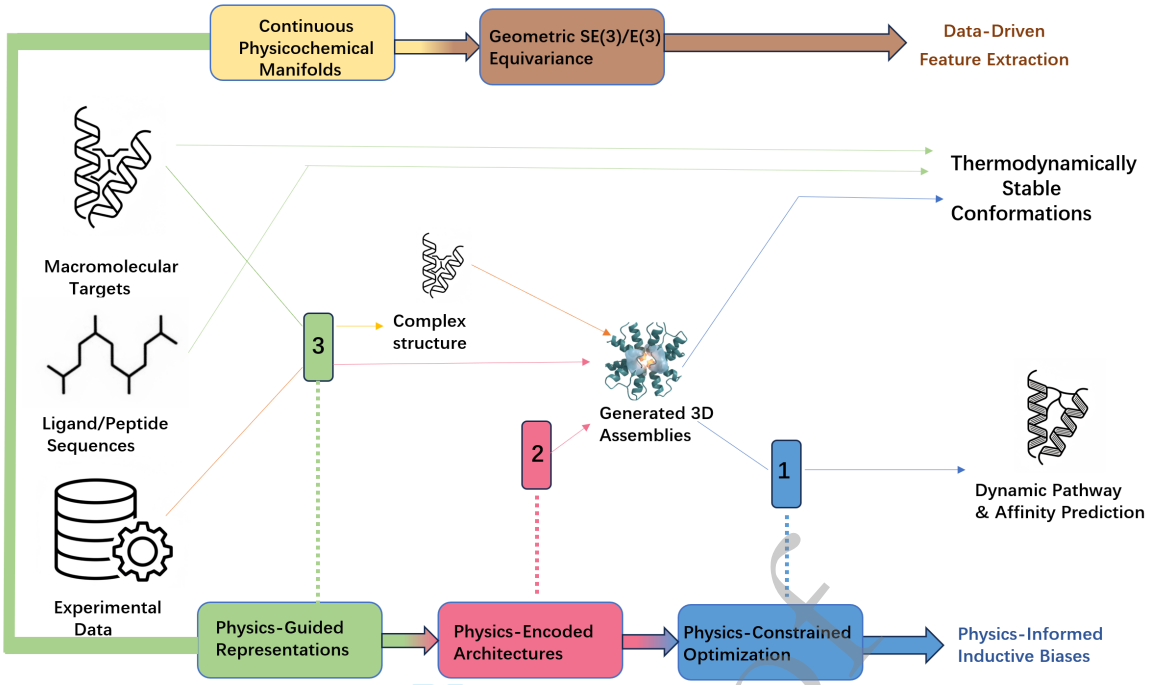


Figure 2. The global framework of Physics-Informed Deep Learning This structural roadmap illustrates the data flow and integration of the three methodological pillars. Raw inputs (left) are processed through a synergistic paradigm of data-driven feature extraction (top) and physics-informed inductive biases (bottom) to yield robust structural predictions (right). The bottom pathway sequentially integrates physics-guided representations at the input level, physics-encoded architectures within the network topology, and physics-constrained optimization at the output level.

While Figure **??** outlines the global data-flow pipeline, to truly operationalize this framework, Figure **??** abstracts the core mathematical and algorithmic mechanisms that define its three pillars. Rather than reviewing isolated models, this mechanistic breakdown provides the foundational paradigmsnavigating from continuous manifold projection at the input level, through geometric equivariance within the network topology, to thermodynamic optimization at the output levelthat will be systematically evaluated in the subsequent sections.

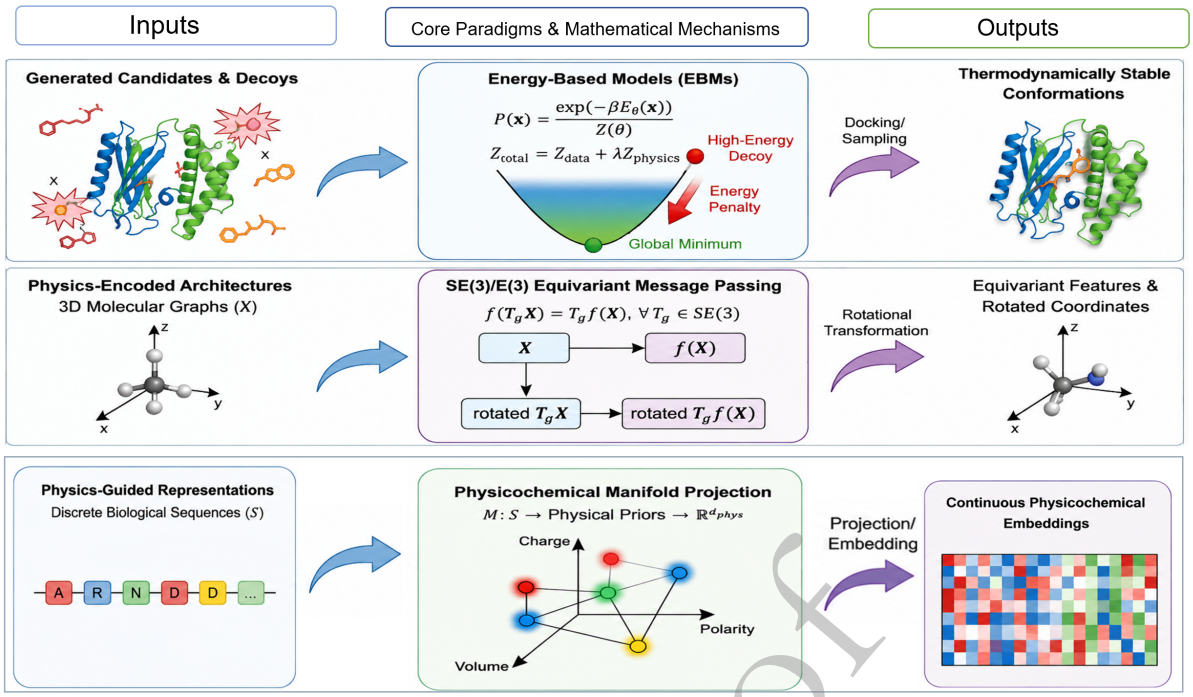


Figure 3. Core algorithmic mechanisms of the Physics-Informed Deep Learning paradigm This diagram illustrates the fundamental input-to-output mathematical transformations of the three methodological pillars. The top panel depicts physics-constrained optimization, where energy-based models minimize composite loss to output thermodynamically stable conformations. The middle panel shows the commutative properties of physics-encoded architectures operating on 3D molecular graphs. The bottom panel illustrates physics-guided representations, projecting discrete biological sequences into continuous physicochemical embeddings.

PHYSICS-CONSTRAINED OPTIMIZATION FOR INTEGRATIVE MODELING OF MACROMOLECULAR ASSEMBLIES

The defining characteristic of Physics-Constrained Optimization is the integration of physical principles directly into the computational model's objective function to enforce thermodynamic consistency.[??] In standard deep learning, models typically optimize empirical metrics such as Root Mean Square Deviation (RMSD)—geometric proxies that frequently lack thermodynamic causality. Consequently, purely data-driven models are prone to generating structures in high-energy states, characterized by steric clashes, solvent-exposed hydrophobic cores, or unrealistic bond geometries. This occurs because such models blindly interpolate the training manifold without perceiving the underlying energy landscape. Physics-constrained frameworks mitigate these limitations by reformulating the learning objective as a composite loss: $\mathscr{L}_{total} = \mathscr{L}_{data} + \lambda \mathscr{L}_{physics}$. This additional term serves as a regularization penalty derived from statistical mechanics, ensuring that the model's predictions are not merely geometrically similar to the ground truth, but strictly converge upon valid local minima on the true free-energy surface. In this section, we analyze two representative paradigms: information-driven docking via explicit

experimental restraints, and statistical mechanical modeling via implicit pseudo-Hamiltonians.

Integrative modeling via explicit restraints

At the explicit constraint level, information-driven integrative modeling addresses the critical "blind search" problem inherent in molecular docking. Conventional *ab initio* algorithms, which rely solely on geometric shape complementarity, frequently become computationally intractable and physically inaccurate for highly flexible systems due to the "induced fit" phenomenon.[?] To circumvent this limitation, the HADDOCK framework establishes a paradigm where sparse, heterogeneous experimental dataranging from NMR chemical shift perturbations to cross-linking mass spectrometryare incorporated directly into the docking force field as thermodynamic penalties.[?,?,?,?,?]

The core theoretical innovation lies in translating ambiguous biological data into a physical force field component known as Ambiguous Interaction Restraints (AIRs).[?,?] Rather than enforcing rigid atom-to-atom distances that would violate the dynamic nature of fluid interfaces, this framework constructs a smooth, funnel-shaped potential energy surface. It converts experimental data into an effective distance potential d_{iAB}:

$$d_{iAB} = \left(\sum_{m \in \{i\}} \sum_{n \in \{j\}} \frac{1}{r_{mn}^{6}} \right)^{-1/6} \tag{1}$$

where the r^{-6} summation acts as an explicit physical prior mimicking the distance-dependence of van der Waals interactions. These AIRs function as "virtual springs" that dramatically reduce the conformational search space. Consequently, the docking challenge is transformed from a random global search into a physically guided optimization process. This strategy successfully drives simulations of flexible systems (e.g., antibody-antigen complexes) toward the native thermodynamic minimum, evidenced by the high structural fidelity and the strong correlation between energy scores and i-RMSD.[?,?,?,?]

Statistical mechanical modeling via energy-based models

While integrative modeling relies on external experimental restraints, a more generalizable frontier involves implicitly embedding thermodynamic principles directly into the neural network architecture through Energy-Based Models (EBMs). In this paradigm, neural networks abandon the traditional classification boundary (probability of 0 or 1) and instead learn to approximate a scalar energy function, $E_\theta(\mathbf{s}_d, \mathbf{s}_p)$, which assigns low energy to stable, native-like conformations and high energy to non-native decoys.[?,?,?,?] Consequently, the probability of a biomolecular interaction is reformulated to strictly ad-

here to the Boltzmann distribution:

$$P(\text{binding} \mid \mathbf{s}_d, \mathbf{s}_p) = \frac{\exp\left(-\beta E_\theta(\mathbf{s}_d, \mathbf{s}_p)\right)}{Z(\theta)} \tag{2}$$

where β is the inverse thermodynamic temperature and $Z(\theta)$ is the intractable partition function.

A representative framework, Hierarchical Statistical Mechanical Modeling (HSM), explicitly parameterizes this energy function as a *pseudo-Hamiltonian*.[?] Rather than treating binding affinity as a monolithic black-box prediction, HSM mathematically reconstructs the global binding energy as the hierarchical sum of localized residue-residue interaction potentials $\Delta\Delta G(a_i, a_j, \theta_{i,j})$:

$$E_\theta(\mathbf{s}_d, \mathbf{s}_p) = \sum_{i \in \text{domain}} \sum_{j \in \text{peptide}} \Delta\Delta G(a_i, a_j, \theta_{i,j}) \tag{3}$$

By employing contrastive divergence or noise-contrastive estimation during training, the loss function acts as a thermodynamic objective that actively "pulls down" the energy of native interactions while "pushing up" the energy of generated decoys. This paradigm theoretically guarantees that the learned latent space mirrors the true multidimensional free-energy landscape, supported by recent advancements in generalizable affinity ranking.[? ? ?] Crucially, this allows models to transfer fundamental interaction principles across different protein families and decode subtle specificity mechanisms (e.g., in SH3 domains) without the need for exhaustive experimental sampling. Furthermore, recent work has begun to incorporate entropic contributions and solvation effects into these scoring functions, paving the way for diffusion-based potentials in generative tasks.[? ? ? ?] Furthermore, it is critical to recognize that optimizing solely for local binding affinitya common pitfall in pure deep learningoften yields broadly adhesive, excessively hydrophobic molecules lacking target specificity. To mitigate this, EBMs and physics-constrained frameworks inherently support *negative design*. By reformulating the objective function to explicitly penalize binding to homologous off-targets ($E_{total} = E_{target} - \lambda E_{off-target}$) or by mathematically constraining the solvent-accessible surface area (SASA) of hydrophobic patches, PIDL actively filters out non-specific "sticky" candidates. This multi-objective physical constraint ensures that the model generates conformations with both high affinity and rigorous selectivity, drastically reducing off-target liabilities.

PHYSICS-ENCODED ARCHITECTURES FOR STRUCTURAL RECOGNITION OF PEPTIDEPROTEIN INTERACTIONS

Unlike physics-constrained optimization which operates at the loss level, Physics-Encoded Architectures embed physical laws directly into the network topology to handle the continuous 3D Euclidean nature of biomolecules. Standard deep learning architectures lack an inherent understanding of 3D continuous spatial symmetries. To ensure biophysical rigor, it is imperative to strictly distinguish between two mathematical groups often conflated in the literature: $SE(3)$ and $E(3)$ equivariance. $SE(3)$ (Special Euclidean) equivariance accounts solely for 3D continuous translations and rotations. Crucially, it preserves molecular chirality (handedness), making it the strict requirement for modeling stereospecific ligand binding. In contrast, $E(3)$ (Euclidean) equivariance encompasses the full Euclidean group by additionally incorporating discrete reflections and inversions. Without embedding these specific symmetries natively into the network topology, models often require inefficient data augmentation and risk generating physically inconsistent predictions. To address this, the paradigm of Geometric Deep Learning (GDL) mathematically bakes these spatial symmetries into the neural message-passing layers, thereby guaranteeing physical consistency by design.[? ?] This shift ensures that the model perceives molecular geometry not as arbitrary data points, but as structured entities governed by fundamental Euclidean laws. In this section, we explore how Equivariant Graph Neural Networks (EGNNs) have enabled generative molecular dynamics, and how advanced $E(3)$-equivariant transformers have revolutionized unified structural modeling for complex biological interfaces.[? ? ?]

Equivariant graph neural networks (EGNN) and generative diffusion

The Equivariant Graph Neural Network (EGNN) fundamentally redefines how neural networks process 3D molecular graphs by strictly ensuring $SE(3)$-equivariance during message passing.[? ? ?] Mathematically, if a transformation $T_g \in SE(3)$ is applied to the input molecule, the predicted coordinate outputs are transformed by the exact same T_g:

$$f(T_g\mathbf{X}) = T_g f(\mathbf{X}) \tag{4}$$

This architecture, supported by foundational work in tensor field networks and SE(3)-transformers, has powered landmark generative frameworks such as DiffDock for protein-ligand docking and RFdiffusion for *de novo* protein design.[? ? ? ?] DiffDock formulates docking as a generative diffusion process over the product space of ligand translations and rotations. By using an $SE(3)$-equivariant score-matching

model, potentially enhanced by high-order force fields or Newtonian message passing, it guarantees that predicted denoising forces rotate in perfect sync with the complex, enabling zero-shot prediction of transient binding poses.[????] Similarly, RFdiffusion leverages the RoseTTAFold architecture to solve the inverse folding problem. It treats the protein backbone as a cloud of residues diffusing in 3D space. By conditioning the denoising process on specific functional motifs, RFdiffusion can "hallucinate" completely novel protein binders that possess high thermodynamic stability, proving that hard-coding spatial symmetries is far more data-efficient than learning them from scratch.[???] To illustrate the practical superiority of this paradigm, consider the real-world therapeutic design using RFdiffusion. In challenging scenarios such as designing *de novo* protein binders against the SARS-CoV-2 spike protein or the interleukin-7 receptor (IL-7Rα)targets where traditional high-throughput screening often failsRFdiffusion achieved unprecedented experimental success rates (up to 20% in wet-lab validation).[?] Because the $SE(3)$-equivariant architecture inherently restricts the generative search space to physically plausible geometries, these computationally “hallucinated” binders required zero subsequent affinity maturation to achieve nanomolar binding affinities, powerfully highlighting how physics-encoded deep learning dramatically accelerates the transition from in silico blueprints to viable therapeutics.

E(3)-equivariant neural networks and unified assembly modeling

Building upon EGNNs, $E(3)$-Equivariant Neural Networks extend physical encoding to encompass the full Euclidean group, including reflections (chirality).[???] This paradigm has catalyzed the development of unified foundation models capable of modeling generalized biomolecular assemblies. A prime example is RoseTTAFold All-Atom (RFAA), which utilizes a novel three-track neural network architecture.[???] Unlike previous sequence-only models, RFAA simultaneously processes 1D sequence information, 2D pairwise distance maps, and 3D coordinate geometry. Its 3D track strictly enforces $SE(3)$-equivariance, allowing it to iteratively refine the stereochemistry of proteins, nucleic acids, small molecules, and metal ions within a single unified graph, effectively bridging the gap between folding and docking.

Following this trajectory, AlphaFold 3 further evolves the paradigm by replacing the MSA-centric Evoformer with a simplified *Pairformer* module and delegating structural generation to a generative Diffusion Module.[??] This module operates on raw atom coordinates, denoising a random cloud into a physically valid structure to model the joint probability of complex assemblies. Parallelly, the Chai-1 architecture pushes the boundaries by addressing the "biological dark matter".[?] Unlike models heavily

reliant on evolutionary co-variation, Chai-1 integrates strictly $E(3)$-equivariant structural modules that allow it to achieve state-of-the-art accuracy even in "single-sequence" mode.[??] Furthermore, it seamlessly extends prediction to diverse chemical spaces, including complex glycans and post-translational modifications (PTMs). Accurately modeling these atypical components requires moving beyond standard 1D amino acid vocabularies. To achieve this, models like Chai-1 and RFAA bypass traditional sequence-only embeddings and natively encode glycans and PTMs as explicit 2D chemical graphs. By embedding exact atom types, bond topologies, and partial charges directly into the network's initial representation, this prior information is seamlessly integrated into the $E(3)$-equivariant structural tracks.[?] This explicitly physical encoding ensures that the neural network can accurately perceive the altered local electrostatic environments and steric hindrances introduced by glycosylation or phosphorylation, fundamentally expanding the model's predictive domain beyond canonical proteins. Complementing these multimodal frameworks, the latest advances in large protein language modeling have catalyzed a paradigm shift toward MSA-free structure prediction. Most notably, models like ESMFold2 leverage the deep evolutionary and biophysical representations distilled by massive language models to directly map single sequences to 3D coordinates. By coupling these rich implicit priors with an equivariant folding head, ESMFold2 achieves high-fidelity structure prediction at unprecedented speeds, demonstrating that language models can effectively internalize the geometric and thermodynamic rules governing macromolecular folding without relying on explicit homologous alignments.[?] By enforcing physical geometry within the network topology, these models demonstrate that physics-encoded architectures can compensate for data sparsity, effectively deducing binding mechanics from fundamental physical constraints.[????]

PHYSICS-GUIDED REPRESENTATIONS FOR SEQUENCE-BASED INTERACTION AND AFFINITY PREDICTION

The paradigm of Physics-Guided Representations integrates physical priors explicitly at the input and feature extraction levels, addressing the fundamental limitations of standard one-hot encoding. Traditional categorical encodings treat biological sequences as discrete strings, thereby ignoring biophysical reality by assigning equidistant values to conservative mutations and radical substitutions alike. To embed physical consistency before neural inference begins, these frameworks project discrete sequence symbols into continuous, low-dimensional physicochemical manifolds. By explicitly incorporating properties such as electrostatic charge, hydrophobicity, secondary structure propensity, and molecular

volume, the input features themselves serve as a definitive physical prior. This strategy ensures that the latent space remains grounded in molecular reality, enhancing the model's ability to generalize across diverse chemical spaces. In this section, we analyze the evolution of these representations: from explicit feature engineering and mathematical manifold projection to the emerging paradigm of implicit biophysical priors distilled by large Protein Language Models (PLMs).

Explicit physicochemical feature engineering in peptide-protein interactions

Peptide-protein interactions (PPIs), which act as critical nodes in cellular signal transduction, are often driven by a duality of localized Short Linear Motifs (SLiMs) and global conformational dynamics.[???] To capture this without the prohibitive cost of MD simulations, computational methods have evolved from basic sequence analysis to sophisticated deep learning architectures.[????] Traditional docking methods and sequence-based predictors often struggle with the flexibility of peptides.[????] Frameworks like CAMP (Convolutional Attention-based Multi-level Prediction) integrate implicit physicochemical properties at the input level using a hybrid architecture.[??] 1D Convolutional Neural Networks (CNNs) function as efficient motif detectors for short-range biochemical complementarity:

$$z_i = f(W \cdot X_{i:i+k-1} + b) \tag{5}$$

simultaneously, Self-Attention mechanisms capture long-range functional coupling between distant residues. A core innovation in CAMP is the use of multi-task learning ($\mathscr{L}_{total} = \mathscr{L}_{interaction} + \lambda \mathscr{L}_{residue\_binding}$). This objective function acts as a spatial regularizer, explicitly forcing the model to identify biologically meaningful binding hotspots rather than overfitting to noise. This mirrors efforts in general PPI prediction and domain-specific modeling (e.g., SH2/SH3 domains).[?????] This physics-guided approach allows the model to accurately identify critical hydrophobic residues in highly flexible therapeutic peptides. For example, in the design of Glucagon-like peptide-1 receptor (GLP-1R) agonists for metabolic diseases, purely sequence-based attention models frequently misclassify the dynamic N-terminus as disordered noise. However, by explicitly encoding molecular volume and charge manifolds, PIDL frameworks can accurately capture the crucial hydrophobic interactions required for receptor activation, correlating strongly with experimental alanine scanning mutagenesis data. This specific structural simulation confirms that explicit physicochemical features serve as a robust proxy for decoding complex thermodynamic binding mechanisms.[?]

Physicochemical manifolds in immune recognition

The need for physical priors is most acute in immune recognition, specifically in the context of cancer immunotherapy and neoantigen identification.[?????] Accurate prediction of MHC ligand binding and subsequent immunogenicity is critical.[?????] However, the theoretical TCR repertoire diversity ($> 10^{15}$) creates a vast "structural dark matter".[???] To overcome the extreme sparsity of training data, models like pMTnet map discrete sequences into a continuous physicochemical manifold using Atchley Factors.[??] These factors condense high-dimensional amino acid indices into five orthogonal physical dimensions (polarity, secondary structure propensity, molecular volume, codon diversity, and electric charge), mapping each residue to a vector $v_i \in \mathbb{R}^5$.

By projecting diverse TCR and peptide sequences into this shared physical space, the neural network learns to identify true physicochemical complementarity (e.g., electrostatic matching between CDR3 loops and epitopes) rather than overfitting to specific sequence k-mers.[?????] This physics-guided representation has proven pivotal in identifying neoantigen immunogenicity and stratifying patient responses to immunotherapy. Deep learning approaches, including TITAN and other TCR-specific models, have further refined this capability.[?????] High-throughput sequencing and single-cell analysis continue to expand our understanding of the TCR repertoire.[??????] These computational insights are directly translatable to clinical settings, aiding in the prediction of responses to immune checkpoint blockade (e.g., PD-1/PD-L1 inhibitors) and understanding mechanisms of resistance.[?????]

Implicit biophysical priors via protein language models (PLMs)

While explicit physicochemical priors offer interpretability, a more powerful paradigm has emerged: *implicit* biophysical priors learned by large Protein Language Models (PLMs) such as ESM-2.[?] Trained on billions of evolutionary sequences via self-supervised masked language modeling, these models effectively "distill" the physical laws governing protein stability and folding. The theoretical justification is that evolution acts as a grand biophysical experiment: sequences that persist through natural selection must satisfy rigorous thermodynamic constraints (i.e., folding into a stable minima with $\Delta G < 0$). Consequently, the embeddings generated by PLMs serve as dynamic, context-aware physicochemical descriptors that surpass static indices. Frameworks like Ankh leverage these pre-trained representations to predict protein-protein interactions with minimal fine-tuning.[?] Unlike static descriptors, PLM embeddings are context-dependentthe "physical vector" of a residue shifts based on its spatial neighbors,

mirroring the cooperative nature of protein folding and marking a shift toward data-driven, evolutionarily informed physical manifolds.

Physics-guided encoding for small molecules and nucleic acids

To fully realize the vision of universal biomolecular modeling, the principles of physics-guided representation must extend beyond amino acid sequences to encompass the vast chemical space of small-molecule ligands and the highly charged, flexible nature of nucleic acids (DNA/RNA).

For small-molecule drug discovery, traditional representations relying on 1D SMILES strings fundamentally lack spatial and biophysical context. Modern physics-informed frameworks overcome this by embedding ligands as 2D molecular graphs or 3D conformer ensembles, explicitly encoding atom types, hybridization states, partial charges, and pharmacophore features as node and edge attributes.[??] By projecting these explicitly physical graphs into the same latent manifold as the receptor pocket, models can learn true physicochemical complementarity (e.g., hydrogen bonding networks and π-π stacking) rather than superficial sequence-string correlations.

Similarly, modeling protein-nucleic acid interactions presents unique physical challenges dominated by the highly anionic phosphate backbone and intricate base-pairing thermodynamics. Purely sequence-based attention models often fail to capture the long-range electrostatic repulsions and the hierarchical folding pathways of RNA. To address this, emerging architectures explicitly integrate RNA secondary structure (e.g., stem-loop topologies and Watson-Crick base-pairing probabilities) as a geometric prior into the neural representation.[??] By utilizing physics-derived base-pairing graphs as an inductive bias, these models can accurately predict 3D ribonucleoprotein assemblies and DNA-transcription factor complexes, demonstrating that structural universality is achieved not by treating all molecules as identical data strings, but by honoring their distinct biophysical constraints at the input level.

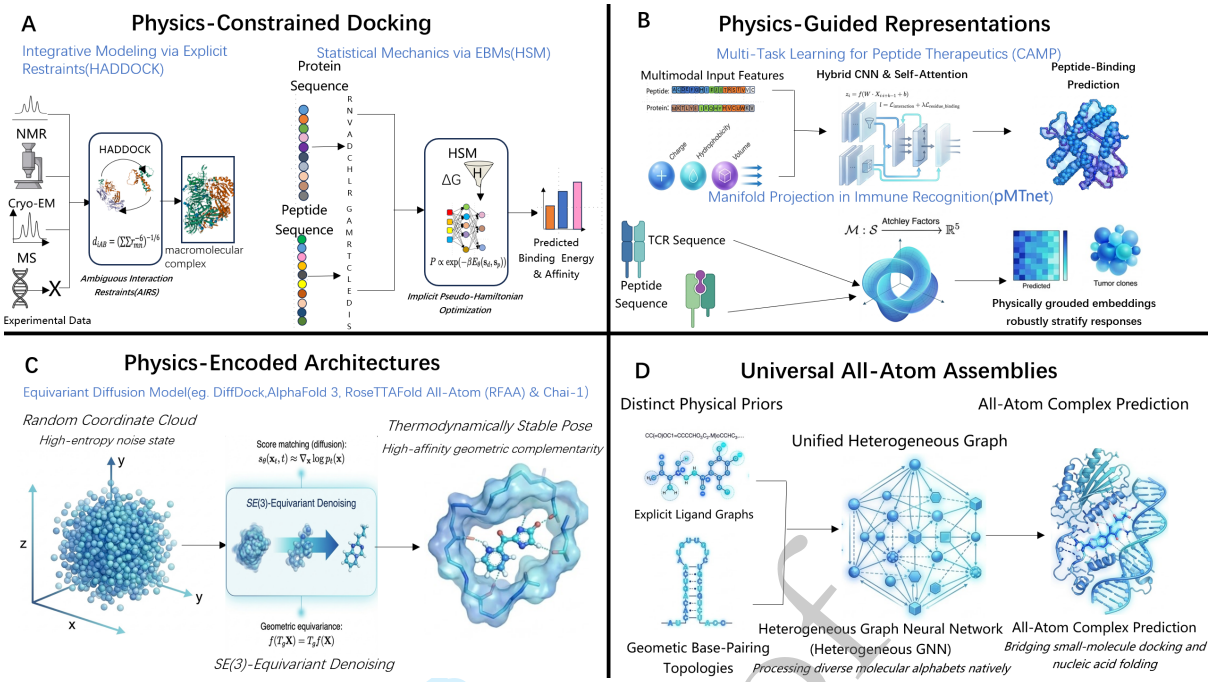


Figure 4. Translational applications of the PIDL paradigm across diverse biological scales and therapeutic modalities This composite plate demonstrates how the methodological pillars are operationalized to solve distinct biomolecular challenges. (A) Physics-constrained docking optimizes flexible assemblies toward true thermodynamic minima. (B) Physics-guided representations project discrete sequences into continuous physicochemical manifolds to identify binding hotspots. (C) Physics-encoded architectures embed spatial symmetries into generative models to resolve high-entropy noise into stable structural poses. (D) Universal all-atom assemblies integrate distinct physical priors across diverse modalities to predict heterogeneous complexes.

Summary of computational resources

To facilitate practical applications and further research, we have summarized the key computational frameworks, tools, and their accessible web servers or open-source repositories discussed in this review (Table **??**). The models listed in the table have been described in the preceding chapters.

Table 1. Summary of Key Computational Models and Resources for Biomolecular Interaction Modeling

Model / Tool	Target System	Core Architecture	Input Features / Priors	Availability
HADDOCK2.4	Macromolecular Complexes	Energy Minimization & Rigid/Flexible Docking	3D Coordinates + Explicit Experimental Restraints	https://wenmr.science.uu.nl/haddock2.4
DiffDock	Protein-Ligand	SE(3) Generative Diffusion	3D Protein Structure + 2D Ligand Graph	https://github.com/gcorso/DiffDock
RFdiffusion	De novo Protein Design	SE(3)-Equivariant Diffusion (RoseTTAFold-based)	Structural Motifs, Contigs, or Pure Noise	https://github.com/RosettaCommons/RFdiffusion
AlphaFold 3	Universal Assemblies	Pairformer & Generative Diffusion Module	Sequences, Ligand SMILES, MSA	https://alphafoldserver.com
RoseTTAFold All-Atom	Universal Assemblies	1D/2D/3D Track Neural Network	Sequences, MSA, Chemical Graphs	GitHub
Chai-1	Universal Assemblies	E(3)-Equivariant Structural Module	Single Sequences, Glycans, PTMs	https://github.com/chaidiscovery/chai-lab
ESM-2	General Purpose Representation	Transformer-based Masked Language Modeling	Raw Sequences (Evolutionary Scale)	https://github.com/facebookresearch/esm
Ankh	General Purpose / Interaction Tasks	Optimized T5 Transformer (Encoder-Decoder)	Raw Sequences (Context-Aware Embeddings)	https://github.com/agemagician/Ankh
CAMP	Peptide-Protein	1D CNN + Self-Attention (Multi-task)	Sequence + Implicit Physicochemical Properties	https://github.com/twopin/CAMP
HSM	Peptide-Protein	Hierarchical Statistical Mechanics	Sequence + Pseudo-Hamiltonian	https://github.com/aqlaboratory/hsm
pMTnet	TCR-pMHC	Transfer Learning + Deep Autoencoder	TCR/Peptide Sequences + Atchley Factors (Manifold)	https://github.com/tianshilu/pMTnet

Abbreviations: MSA, multiple sequence alignment; SMILES, simplified molecular-input line-entry system; PTMs, post-translational modifications; CNN, convolutional neural network; HSM, hierarchical statistical mechanical modeling; TCR, T-cell receptor; pMHC, peptide-major histocompatibility complex.

While specific computational tools have been summarized previously (Table **??**), selecting an appropriate PIDL paradigm requires a more systematic methodological evaluation. To this end, Table **??**

provides a comprehensive multidimensional comparison of the three core PIDL categories (Physics-inspired, Physics-informed, and Physics-encoded). This comparison systematically evaluates their respective physical priors, constraint rigidity (soft vs. hard), applicable biomolecular tasks, interpretability, computational costs, and inherent limitations. This matrix serves as a methodological guide for selecting the optimal algorithmic architecture based on the specific thermodynamic or geometric demands of the target system.

Table 2. Systematic Comparison of Physics-Informed Deep Learning (PIDL) Categories. This table evaluates the three distinct methodologies based on their physical constraints, applicability, computational trade-offs, and validation extent.

PIDL Category	Physical Prior & Constraint Type	Inputs → Outputs	Applicable Tasks & Systems	Interpretability	Computational Cost	Major Advantages	Key Limitations
Physics-Inspired (e.g., Generative Diffusion)	Non-equilibrium thermodynamics (*Soft/No constraint*)	Random noise + conditions → 3D coordinates	*De novo* generation, conformer sampling (Proteins, Ligands)	**Low:** Black-box stochastic denoising trajectory.	**High:** Requires thousands of iterative denoising steps.	Unprecedented structural diversity; ability to explore novel chemical spaces.	Prone to thermodynamic "hallucinations"; lacks explicit energy minima guarantees.
Physics-Informed (e.g., EBMs, Restrained Docking)	Boltzmann distributions, Force fields (*Soft constraint via Loss*)	3D poses/sequences → Energy scores (ΔG)	Affinity prediction, complex docking, pose ranking	**High:** Explicit energy terms and identifiable biophysical interactions.	**Moderate:** Efficient scoring, but relies on prior sampling.	Strictly guarantees local thermodynamic stability; easily integrates experimental data.	Primarily discriminative; struggles with *de novo* generation from scratch.
Physics-Encoded (e.g., $SE(3)/E(3)$ Equivariant Networks)	Euclidean spatial symmetries (*Hard structural constraint*)	Molecular graphs → Equivariant features, Coordinates	Structure prediction, rigid-body docking (All-atom assemblies)	**Medium:** Preserves geometric validity, but features remain latent.	**Moderate-High:** Complex tensor operations scale quadratically.	Absolute geometric consistency; highly data-efficient without data augmentation.	Architecturally rigid; difficult to model highly disordered or high-entropy regions.

CRITICAL ASSESSMENT: THE SYSTEMIC MISMATCH

Despite the transformative impact of Physics-Informed Deep Learning (PIDL), the field has reached a conceptual plateau where incremental performance gains often mask systemic mismatches between statistical models and biological reality. Current paradigms remain largely reductionist, modeling macromolecules as static entities in a thermodynamic vacuum and relying on sequence-centric attention mechanisms that often struggle to capture the multi-scale, dynamic, and ultra-long nature of cellular life. While benchmarks show high accuracy on rigid targets, these models struggle to internalize the governing laws of statistical mechanics, leading to predictions that are geometrically consistent yet thermodynamically improbable. In this section, we critically interrogate these limitations through four systemic bottlenecks: the generalizability crisis driven by crystallographic bias, the emergence of structural hallu-

cinations from entropic blindness, the mathematical failure of current architectures to encode full-scale interactomes due to sequence truncation, and the persistent operational divide between discriminative scoring and autonomous generative discovery.

The generalizability crisis and crystallographic bias

The predictive power of modern foundation models is frequently a statistical artifact of overfitting to the structural distributions inherent in the Protein Data Bank (PDB). This "crystallographic bias" stems from the fact that available training data overwhelmingly represent stable, low-entropy conformations optimized for experimental tractability (e.g., crystallization or Cryo-EM reconstruction).?? Consequently, neural networks internalize a distorted view of protein physics, equating structural validity with rigidity. This limits their ability to generalize to "biological dark matter," such as Intrinsically Disordered Proteins (IDPs), transient signaling intermediates, or metamorphic folds, where structural heterogeneity is functionally decisive.?? Furthermore, the generalizability crisis is severely exacerbated by the widespread problem of benchmark overfitting. Most state-of-the-art AI docking models (such as DiffDock) are highly optimized and evaluated on standard datasets like PDBBind. However, the ligand molecular structures and pocket rigidities within these training sets are highly homogeneous. As recent critical evaluations have demonstrated, when these deep learning models are confronted with entirely new chemical spaces or highly flexible targets that lack homologous evolutionary information, their predictive accuracy and physical consistency plummet sharply. This over-reliance on homogeneous benchmarks creates an illusion of progress, highlighting that current models often learn superficial structural biases rather than true, generalizable physical interactions.Unlike physics-based MD simulations that can explore non-equilibrium states, static learning frameworks lack the capacity to reason over conformational ensembles, leading to significant performance degradation when identifying targets outside the "stable fold" manifold.

AI "hallucinations" and entropic blindness

A striking failure of current generative models is the production of "hallucinations"structures that appear geometrically plausible under metrics like pLDDT but are thermodynamically highly unfavorable in solution. This phenomenon arises from a fundamental *entropic blindness*: while deep models are adept at capturing local enthalpic regularities (e.g., hydrogen bonding networks and secondary structures), they systematically underestimate the massive entropic penalties associated with solvent desolvation, loop

flexibility, and hydrophobic exposure.[?,?] In the absence of an explicit free-energy objective function (ΔG), latent space optimizations often converge on non-physical manifolds, producing configurations with severe steric clashes or solvent-exposed hydrophobic cores that violate the repulsive terms of the Lennard-Jones potential. Fundamentally, generative models approximate the data distribution $P_{data}(x)$, which is distinct from the physical Boltzmann distribution $P_{phys}(x) \propto e^{-E(x)/k_BT}$. Without anchoring predictions to the true thermodynamic partition function, these hallucinations represent statistical probabilities rather than physical realities, severely limiting their utility in drug design where stability is paramount.

The complexity bottleneck: sequence truncation and allosteric dilution

A critical technical frontier remains unaddressed: the systemic failure of current Protein Language Models (PLMs) to effectively encode ultra-long protein targets and massive multi-chain interactomes. Due to the quadratic computational complexity ($O(N^2)$) of standard Transformer attention mechanisms, researchers are frequently forced to utilize arbitrary sequence truncation, typically capping inputs at 1,024 or 2,048 residues. In the context of large-scale complexessuch as cell-surface receptor clusters, sarcomeric giants like titin, or multi-domain scaffolding proteinsthis practice creates an "information graveyard" at the molecular interface. Critical long-range allosteric signals, which coordinate binding events across distant domains through subtle conformational waves, are mathematically diluted by the overwhelming background of inert residues or simply discarded by truncation. This reliance on 1D sequence-centric encoding effectively treats the molecule as a linear string, failing to perceive the global 3D topology and rendering the model blind to the synergistic specificity required for high-fidelity biologic discovery. Consequently, the model loses the "global context" necessary to understand how a mutation at the N-terminus might structurally destabilize a binding pocket at the C-terminus.

The operational divide: discriminative vs. generative design

There exists a fundamental disconnect between *evaluating* an interaction and *originating* one from first principles. Current state-of-the-art frameworks, such as HADDOCK or HSM, are fundamentally discriminative: they are exceptionally robust at ranking, scoring, or refining candidate interactions supplied to them.[?,?,?] However, they lack the differentiable pathway needed to navigate the astronomical chemical space for *de novo* design. Most discriminative models cannot provide the constructive gradients required to guide generative sampling toward physically stable optima; optimizing against a discrim-

inative score often leads to "adversarial examples"molecules that trick the scorer but are biologically invalid. This "dead-end" logic restricts AI to a passive screening role rather than an active designer, creating a bottleneck where models can identify "what binds" among known candidates but cannot independently deduce "what should be built" to modulate a novel therapeutic target.[?]

FUTURE PERSPECTIVES: ONTOLOGICAL RESTRUCTURING

The next theoretical evolution in biomolecular modeling requires a paradigm shift from incremental architectural scaling to a systematic redefinition of modeling objectives that aligns with the dynamic, context-dependent reality of living systems. We envision this evolution unfolding through four tightly coupled directions that directly address the limitations identified above: the shift toward context-aware mesoscale modeling, the leap from static structures to 4D dynamic trajectories, the adoption of vision-inspired architectures for full-scale encoding and zero-shot physical probing, and the realization of self-evolving discovery engines through autonomous experimental feedback.

Near-term optimization priorities: benchmarks and closed-loop validation

Before fully realizing the visionary goals of autonomous laboratories and 4D simulations, the field must address several immediate implementation strategies to bridge the current translational gap. First, there is an urgent need to establish standardized benchmarks for physical plausibility. Current evaluation metrics relying predominantly on geometric RMSD must be augmented with strictly biophysical scoring functions that heavily penalize steric clashes, unphysical bond geometries, and thermodynamic instabilities. Second, near-term algorithmic development should prioritize integrating Molecular Dynamics (MD)-derived geometry with deep learning. Rather than replacing MD entirely, deep learning can be utilized to rapidly generate initial conformational ensembles, followed by short, machine-learning-accelerated MD relaxation steps to ensure local thermodynamic validity. Finally, computational predictions must be tightly coupled with empirical reality through closed-loop experimental validation. Utilizing high-resolution techniques such as Cryo-Electron Microscopy (Cryo-EM) and Nuclear Magnetic Resonance (NMR) for structural verification, alongside Surface Plasmon Resonance (SPR) for kinetic binding affinity (k_{on}/k_{off}) validation, will provide the essential ground-truth feedback required to continuously recalibrate and refine the physical priors within deep learning models.

From isolated structures to context-aware mesoscale interactomes

To overcome the limitations of crystallographic bias (Section 5.1), structural AI must move beyond "in vacuo" predictions of isolated complexes. The future lies in *Context-Aware Modeling*, which explicitly integrates Coarse-Grained (CG) physics with deep representations to account for the mesoscale cellular environment.[?] This includes modeling the effects of explicit lipid membrane curvature, pH gradients, electrostatic screening, and competitive molecular crowding. By modeling the peptide-protein interactome within its native physiological context rather than as a floating rigid body, researchers can evaluate true *in vivo* efficacy and off-target liability long before clinical trials. This holistic approach will transform structural biology from a static atlas of molecules into a predictive simulator of the cellular state, enabling the design of drugs that utilize environmental factors (e.g., pH-dependent binding) as functional mechanisms.

Beyond static minima: 4D spatiotemporal pathway generation

To resolve the problem of entropic blindness (Section 5.2), the field must advance from seeking static free-energy minima to generating continuous dynamical trajectories. We propose the development of *4D Spatiotemporal Generative Models* that utilize Physics-Informed Neural Operators (PINOs) to solve non-equilibrium Langevin equations.[? ?] Instead of predicting a single final bound pose, future models will learn to generate the entire kinetic pathway of a binding event, encompassing transient intermediates, cryptic pockets, and high-energy transition states. This capability will allow AI to quantify critical dynamic properties such as drug residence time (k_{off}) and association rates (k_{on}), which often correlate better with clinical efficacy than simple affinity (K_D). This shifts the goal of molecular AI from predicting the thermodynamic "destination" to simulating the complete physical "journey."

Vision-inspired architectures and zero-shot physical probing

To bypass the inherent length constraints and complexity bottlenecks of 1D PLMs (Section 5.3), we advocate for a paradigm shift toward *Multi-scale Geometric Architectures*. Rather than relying on inappropriate analogies to 2D image processing or document scanning, macromolecules must be modeled natively as high-dimensional geometric manifolds or continuous electron density fields. This direction draws inspiration from Hierarchical Coarse-Graining (CG) Graph Neural Networks and Sparse 3D Equivariant Convolutions. By dynamically pooling atomic representations into residue-level, and subsequently domain-level graphs, or by treating the molecular surface as a voxelized 3D scalar field (akin to

high-resolution Cryo-EM density maps), this approach circumvents the $O(N^2)$ sequence limit entirely, preserving the intricate allosteric signals necessary for capturing global functional specificity. Crucially, the rich internal representations of these advanced architectures can be repurposed for *zero-shot physical probing*. Inspired by frameworks like AF2BIND[?] which extracts pretrained pair representations from AlphaFold2 to probe small-molecule pockets using disconnected amino acid "baits"future models can deploy a similar strategy for peptide-protein interactions. By feeding structured peptide motifs (e.g., short α-helices or randomized k-mers) as geometric probes into these full-scale networks, researchers can bypass massive explicit training. Instead, they can directly extract latent attention matrices to map the continuous binding-energy landscape of the receptor surface, effectively transforming foundation models from static structure predictors into dynamic, zero-shot biophysical sensors.

Autonomous molecular discovery via self-driving laboratories

A compelling strategy to bridge the discriminative-generative divide (Section 5.4) is the integration of AI with autonomous robotics in a "Lab-in-the-Loop" configuration. In this paradigm, Large Language Model (LLM) scientific agents orchestrate the entire hypothesis-design-synthesis cycle. Unlike static models, these agents use real-time experimental data (e.g., from high-throughput Cryo-EM or SPR assays) to iteratively correct the AI's internal energy landscape priors. Crucially, this system leverages "negative data"failed binding experimentsas hard physical constraints to prune the hypothesis space. By continuously updating its internal beliefs based on physical feedback, the AI system evolves from a fixed predictor into a self-correcting discovery engine. This transition will transform drug design from a stochastic search process into an autonomous, experimentally grounded molecular foundry, capable of navigating the vast chemical space with increasing precision and reduced experimental cost.

In conclusion, this review argues that the future of molecular recognition modeling lies in an integrative paradigm in which data are not used to replace physical models, but to constrain, calibrate, and activate them. By embedding physical consistency into deep learning architectures and aligning computational predictions with experimental realities, the field has the potential to bridge the long-standing gap between structural insight and therapeutic application.

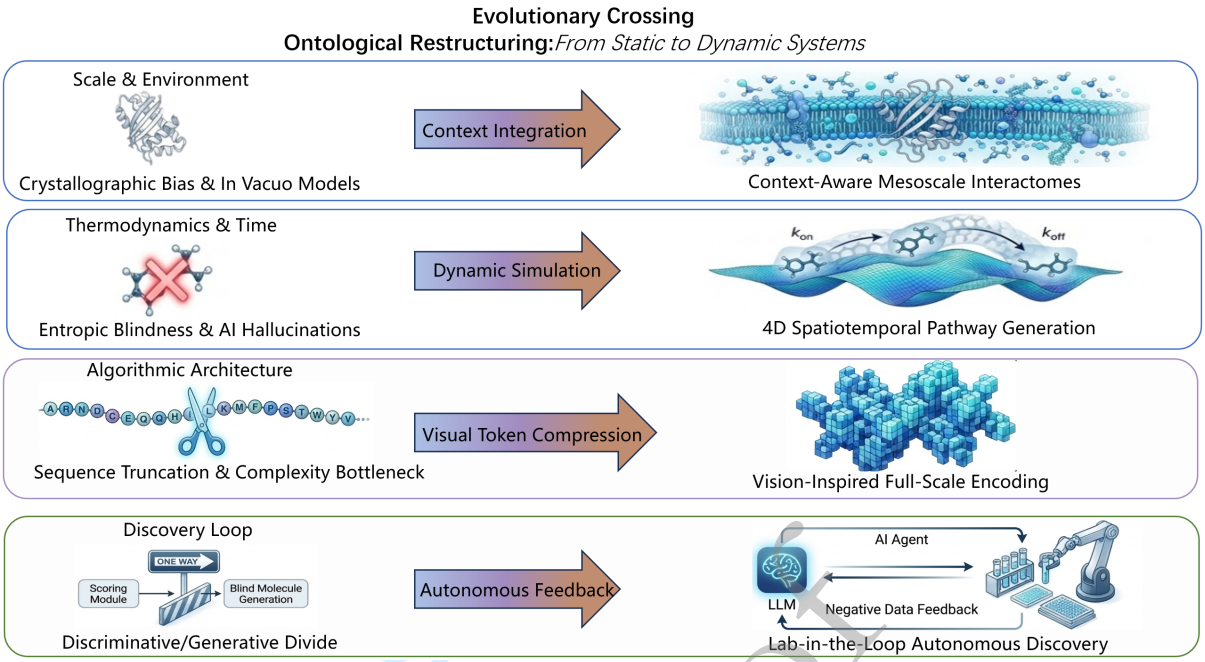


Figure 5. Ontological restructuring: Overcoming systemic mismatches and mapping the future of biomolecular AI This conceptual diagram outlines the necessary evolutionary crossing from current structural bottlenecks (Left) to next-generation modeling paradigms (Right). Row 1 (Scale & Environment): Moving toward context-aware mesoscale interactomes. Row 2 (Thermodynamics & Time): Transitioning from static minima to 4D spatiotemporal pathway generation. Row 3 (Algorithmic Architecture): Resolving sequence truncation via vision-inspired token compression. Row 4 (Discovery Loop): Bridging the discriminative-generative divide through a self-correcting Lab-in-the-Loop engine.

REFERENCES

1. Bohacek R. S., McMartin C. and Guida W. C. (1996). The art and practice of structure-based drug design: a molecular modeling perspective. *Med. Res. Rev.* **16**:3–50. DOI:10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6

2. Reymond J.-L. and Awale M. (2012). Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem. Neurosci.* **3**:649–657. DOI:10.1021/cn3000422

3. Sun D., Gao W., Hu H. et al. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharm. Sin. B* **12**:3049–3062. DOI:10.1016/j.apsb.2022.02.002

4. Oprea T. I., Bologa C. G., Brunak S. et al. (2018). Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**:317–332. DOI:10.1038/nrd.2018.14

5. Lee C. H., Huh J., Buckley P. R. et al. (2023). A robust deep learning workflow to predict $CD8^+$ T-cell epitopes. *Genome Med.* **15**:70. DOI:10.1186/s13073-023-01225-z

6. De Vivo M., Masetti M., Bottegoni G. et al. (2016). Role of molecular dynamics and related methods in drug discovery. *J. Med. Chem.* **59**:4035–4061. DOI:10.1021/acs.jmedchem.5b01684

7. Claveria-Gimeno R., Vega S., Abian O. et al. (2017). A look at ligand binding thermodynamics in drug discovery. *Expert Opin. Drug Discov.* **12**:363–377. DOI:10.1080/17460441.2017.1297418

8. Teague S. J. (2003). Implications of protein flexibility for drug discovery. *Nat. Rev. Drug Discov.* **2**:527–541. DOI:10.1038/nrd1129

9. Lu T., Zhang Z., Zhu J. et al. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nat. Mach. Intell.* **3**:864–875. DOI:10.1038/s42256-021-00383-2

10. DiMasi J. A., Grabowski H. G. and Hansen R. W. (2016). Innovation in the pharmaceutical industry: new estimates of R&D costs. *J. Health Econ.* **47**:20–33. DOI:10.1016/j.jhealeco.2016.01.012

11. Arstila T. P., Casrouge A., Baron V. et al. (1999). A direct estimate of the human $\alpha\beta$ T cell receptor diversity. *Science* **286**:958–961. DOI:10.1126/science.286.5441.958

12. Dudek A. Z., Arodz T. and Gálvez J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High Throughput Screen.* **9**:213–228. DOI:10.2174/138620706776055539

13. Nantasenamat C. (2009). A practical overview of quantitative structure activity relationship. *EXCLI J.* **8**:74–88. DOI:10.17877/DE290R-690

14. Cherkasov A., Muratov E. N., Fourches D. et al. (2014). QSAR modeling: where have you been? Where are you going to? *J. Med. Chem.* **57**:4977–5010. DOI:10.1021/jm4004285

15. Keifer P. A. (2003). Flow injection analysis NMR (FIA–NMR): a novel flow NMR technique that complements LC–NMR and direct injection NMR (DI–NMR). *Magn. Reson. Chem.* **41**:509–516. DOI:10.1002/mrc.1206

16. Morris C. J. and Della Corte D. (2021). Using molecular docking and molecular dynamics to investigate protein–ligand interactions. *Mod. Phys. Lett. B* **35**:2130002. DOI:10.1142/S0217984921300027

17. Hollingsworth S. A. and Dror R. O. (2018). Molecular dynamics simulation for all. *Neuron* **99**:1129–1143. DOI:10.1016/j.neuron.2018.08.011

18. Morris G. M. and Lim-Wilby M. (2008). Molecular docking. *Methods Mol. Biol.* **443**:365–382. DOI:10.1007/978-1-59745-177-2_19

19. Li J., Fu A. and Zhang L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci.* **11**:320–328. DOI:10.1007/s12539-019-00327-w

20. Trott O. and Olson A. J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**:455–461. DOI:10.1002/jcc.21334

21. Wang L., Wu Y., Deng Y. et al. (2015). Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* **137**:2695–2703. DOI:10.1021/ja512751q

22. Gilson M. K. and Zhou H.-X. (2007). Calculation of protein-ligand binding affinities. *Annu. Rev. Biophys. Biomol. Struct.* **36**:21–42. DOI:10.1146/annurev.biophys.36.040306.132550

23. Dror R. O., Dirks R. M., Grossman J. P. et al. (2012). Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.* **41**:429–452. DOI:10.1146/annurev-biophys-042910-155245

24. Perez J. J., Perez R. A. and Perez A. (2021). Computational modeling as a tool to investigate PPI: from drug design to tissue engineering. *Front. Mol. Biosci.* **8**:681617. DOI:10.3389/fmolb.2021.681617

25. Sousa S. F., Fernandes P. A. and Ramos M. J. (2006). Protein–ligand docking: current status and future challenges. *Proteins* **65**:15–26. DOI:10.1002/prot.21082

26. Mohan V., Gibbs A. C., Cummings M. D. et al. (2005). Docking: successes and challenges. *Curr. Pharm. Des.* **11**:323–333. DOI:10.2174/1381612053382106

27. Yang K., Swanson K., Jin W. et al. (2019). Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**:3370–3388. DOI:10.1021/acs.jcim.9b00237

28. Altae-Tran H., Ramsundar B., Pappu A. S. et al. (2017). Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**:283–293. DOI:10.1021/acscentsci.6b00367

29. Ching T., Himmelstein D. S., Beaulieu-Jones B. K. et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15**:20170387. DOI:10.1098/rsif.2017.0387

30. Rossjohn J., Gras S., Miles J. J. et al. (2015). T cell antigen receptor recognition of antigen-presenting molecules. *Annu. Rev. Immunol.* **33**:169–200. DOI:10.1146/annurev-immunol-032414-112334

31. Knapp B., Demharter S., Esmaielbeiki R. et al. (2015). Current status and future challenges in T-cell receptor/peptide/MHC molecular dynamics simulations. *Brief. Bioinform.* **16**:1035–1044. DOI:10.1093/bib/bbv005

32. Chmiela S. et al. (2017). Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**:e1603015. DOI:10.1126/sciadv.1603015

33. Unke O. T. et al. (2021). Machine learning force fields. *Chem. Rev.* **121**:10142–10186. DOI:10.1021/acs.chemrev.0c01111

34. Jiménez-Luna J., Grisoni F. and Schneider G. (2020). Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2**:573–584. DOI:10.1038/s42256-020-00236-4

35. Karniadakis G. E. et al. (2021). Physics-informed machine learning. *Nat. Rev. Phys.* **3**:422–440. DOI:10.1038/s42254-021-00314-5

36. Miyazawa S. and Jernigan R. L. (1996). Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term. *J. Mol. Biol.* **256**:623–644. DOI:10.1006/jmbi.1996.0114

37. Wang Y., Li Y., Chen J. et al. (2025). Modeling protein–ligand interactions for drug discovery in the era of deep learning. *Chem. Soc. Rev.* **54**. DOI:10.1039/d5cs00415b

38. Lei Y. et al. (2021). A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat. Commun.* **12**:5465. DOI:10.1038/s41467-021-25772-4

39. Moon S., Zhung W., Yang S. et al. (2022). PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem. Sci.* **13**:3661–3673. DOI:10.1039/d1sc06946b

40. Deshpande A. (2025). Learning biomolecular motion: the physics-informed machine learning paradigm. *arXiv*:2511.06585.

41. Ugurlu S. Y. et al. (2024). Cobdock: an accurate and practical machine learning-based consensus blind docking method. *J. Cheminform.* **16**:5. DOI:10.1186/s13321-023-00793-x

42. Honorato R. V. et al. (2024). The HADDOCK2.4 web server for integrative modeling of biomolecular complexes. *Nat. Protoc.* **19**:33. DOI:10.1038/s41596-024-01011-0

43. Dominguez C., Boelens R. and Bonvin A. M. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**:1731–1737. DOI:10.1021/ja026939x

44. Koukos P. I., Réau M. and Bonvin A. M. (2021). Shape-restrained modeling of protein–small-molecule complexes with high ambiguity driven DOCKing. *J. Chem. Inf. Model.* **61**:4807–4818. DOI:10.1021/acs.jcim.1c00796

45. Nilges M. (1995). Calculation of protein structures with ambiguous distance restraints: automated assignment of ambiguous NOE crosspeaks and disulphide connectivities. *J. Mol. Biol.* **245**:645–660. DOI:10.1006/jmbi.1994.0053

46. Rout M. P. and Sali A. (2019). Principles for integrative structural biology studies. *Cell* **177**:1384–1403. DOI:10.1016/j.cell.2019.05.016

47. Vajda S. and Kozakov D. (2009). Convergence and combination of methods in protein–protein docking. *Curr. Opin. Struct. Biol.* **19**:164–170. DOI:10.1016/j.sbi.2009.02.008

48. Rodrigues J. P. and Bonvin A. M. (2014). Integrative computational modeling of protein interactions. *FEBS J.* **281**:1988–2003. DOI:10.1111/febs.12771

49. Bonvin A. M. (2006). Flexible protein–protein docking. *Curr. Opin. Struct. Biol.* **16**:194–200. DOI:10.1016/j.sbi.2006.02.002

50. Rodrigues J. P. G. L. M., Karaca E. and Bonvin A. M. J. J. (2014). Information-driven structural modelling of protein–protein interactions. In: Kukol A. (ed.), *Molecular Modeling of Proteins* (Springer), pp. 399–424. DOI:10.1007/978-1-4939-1465-4_18

51. Ward A. B., Sali A. and Wilson I. A. (2013). Integrative structural biology. *Science* **339**:913–915. DOI:10.1126/science.1228565

52. Yin R. and Pierce B. G. (2024). Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy. *Protein Sci.* **33**:e4865. DOI:10.1002/pro.4865

53. Noé F., Olsson S., Köhler J. et al. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**:eaaw1147. DOI:10.1126/science.aaw1147

54. Du Y. and Mordatch I. (2019). Implicit generation and modeling with energy based models. *Adv. Neural Inf. Process. Syst.* **32**.

55. LeCun Y., Chopra S., Hadsell R. et al. (2006). A tutorial on energy-based learning. *Predicting Structured Data* **1**.

56. Zhou H. and Zhou Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* **11**:2714–2726. DOI:10.1110/ps.0217002

57. Cunningham J. M., Koytiger G., Sorger P. K. et al. (2020). Biophysical prediction of protein–peptide interactions and signaling networks using machine learning. *Nat. Methods* **17**:175–183. DOI:10.1038/s41592-019-0687-1

58. Yan J. et al. (2024). Deltadock: a unified framework for accurate, efficient, and physically reliable molecular docking. *Adv. Neural Inf. Process. Syst.* **37**:113282–113310.

59. Li Y. et al. (2025). Decoding the limits of deep learning in molecular docking for drug discovery. *Chem. Sci.* **16**:17374–17390. DOI:10.1039/d5sc05395a

60. Brown B. P. (2025). A generalizable deep learning framework for structure-based protein–ligand affinity ranking. *Proc. Natl. Acad. Sci. U.S.A.* **122**:e2508998122. DOI:10.1073/pnas.2508998122

61. Chang C.-e. A., Chen W. and Gilson M. K. (2007). Ligand configurational entropy and protein binding. *Proc. Natl. Acad. Sci. U.S.A.* **104**:1534–1539. DOI:10.1073/pnas.0610494104

62. Chodera J. D. and Mobley D. L. (2013). Entropy-enthalpy compensation: role and ramifications in biomolecular ligand recognition and design. *Annu. Rev. Biophys.* **42**:121–142. DOI:10.1146/annurev-biophys-083012-130318

63. Wu H. et al. (2024). in *12th Int. Conf. Learn. Represent.*

64. Li J., Beaudoin C. and Ghosh S. (2023). Energy-based generative models for target-specific drug discovery. *Front. Mol. Med.* **3**:1160877. DOI:10.3389/fmmed.2023.1160877

65. Atz K., Grisoni F. and Schneider G. (2021). Geometric deep learning on molecular representations. *Nat. Mach. Intell.* **3**:1023–1032. DOI:10.1038/s42256-021-00418-8

66. Bronstein M. M. et al. (2021). Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv*:2104.13478.

67. Batzner S. et al. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**:2453. DOI:10.1038/s41467-022-29939-5

68. Satorras V. G., Hoogeboom E. and Welling M. (2021). E(n) equivariant graph neural networks. In: *Proc. 38th Int. Conf. Mach. Learn.*, pp. 9323–9332.

69. Schütt K. T., Unke O. T. and Gastegger M. (2021). Equivariant message passing for the prediction of tensorial properties and molecular spectra. In: *Proc. 38th Int. Conf. Mach. Learn.*, pp. 9377–9388.

70. Stärk H., Ganea O.-E., Pattanaik L. et al. (2022). EquiBind: geometric deep learning for drug binding structure prediction. In: *Proc. 39th Int. Conf. Mach. Learn.*, pp. 20503–20521.

71. Ganea O.-E. et al. (2021). Independent SE(3)-equivariant models for end-to-end rigid protein docking. *arXiv*:2111.07786.

72. Hoogeboom E., Satorras V. G., Vignac C. et al. (2022). Equivariant diffusion for molecule generation in 3D. In: *Proc. 39th Int. Conf. Mach. Learn.*, pp. 8867–8887.

73. Thomas N. et al. (2018). Tensor field networks: rotation- and translation-equivariant neural networks. *arXiv*:1802.08219.

74. Fuchs F. B. et al. (2020). SE(3)-transformers: 3D roto-translation equivariant attention networks. *Adv. Neural Inf. Process. Syst.* **33**:1970–1981.

75. Corso G., Stärk H., Jing B. et al. (2022). DiffDock: diffusion steps, twists, and turns for molecular docking. *arXiv*:2210.01776.

76. Watson J. L. et al. (2023). De novo design of protein structure and function with RFdiffusion. *Nature* **620**:1089–1100. DOI:10.1038/s41586-023-06415-8

77. Batatia I., Kovacs D. P., Simm G. et al. (2022). MACE: higher order equivariant message passing neural networks. *Adv. Neural Inf. Process. Syst.* **35**:11423–11436.

78. Musaelian A. et al. (2023). Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**:579. DOI:10.1038/s41467-023-36329-y

79. Haghighatlari M. et al. (2022). NewtonNet: a Newtonian message passing network for deep learning of interatomic potentials and forces. *Digit. Discov.* **1**:333–343. DOI:10.1039/d2dd00008c

80. Schütt K. T. et al. (2017). SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **30**.

81. Qiao Z., Nie W., Vahdat A. et al. (2024). State-specific protein–ligand complex structure prediction. *Nat. Mach. Intell.* **6**:195–208. DOI:10.1038/s42256-024-00792-z

82. Lu W. et al. (2024). DynamicBind: predicting ligand-specific protein-ligand complex structure with a deep equivariant generative model. *Nat. Commun.* **15**:1071. DOI:10.1038/s41467-024-45461-2

83. Duval A. et al. (2023). A hitchhiker's guide to geometric GNNs for 3D atomic systems. *arXiv*:2312.07511.

84. Geiger M. and Smidt T. (2022). e3nn: Euclidean neural networks. *arXiv*:2207.09453.

85. Jing B., Eismann S., Suriana P. N. et al. (2020). Learning from protein structure with geometric vector perceptrons. *arXiv*:2009.01411.

86. Ganea O.-E. et al. (2021). GeoMol: torsional geometric generation of molecular 3D conformer ensembles. *Adv. Neural Inf. Process. Syst.* **34**:13757–13769.

87. Baek M. et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**:871–876. DOI:10.1126/science.abj8754

88. Krishna R. et al. (2024). Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**:eadl2528. DOI:10.1126/science.adl2528

89. Baek M. et al. (2024). Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**:117–121. DOI:10.1038/s41592-023-02086-5

90. Jumper J. et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589. DOI:10.1038/s41586-021-03819-2

91. Abramson J. et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**:493–500. DOI:10.1038/s41586-024-07487-w

92. Chai Discovery. (2024). Chai-1: decoding the molecular interactions of life. *bioRxiv*:2024.10.10.615955. DOI:10.1101/2024.10.10.615955

93. Ovchinnikov S. et al. (2017). Protein structure determination using metagenome sequence data. *Science* **355**:294–298. DOI:10.1126/science.aah4043

94. Hopf T. A. et al. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* **3**:e03430. DOI:10.7554/eLife.03430

95. Grisafi A., Wilkins D. M., Csányi G. et al. (2018). Symmetry-adapted machine learning for tensorial properties of atomistic systems. *Phys. Rev. Lett.* **120**:036002. DOI:10.1103/PhysRevLett.120.036002

96. Zhou G. et al. (2023). Uni-Mol: a universal 3D molecular representation learning framework. In: *11th Int. Conf. Learn. Represent.*.

97. Pei Q. et al. (2023). FABind: fast and accurate protein-ligand binding. *Adv. Neural Inf. Process. Syst.* **36**:55963–55980.

98. AlQuraishi M. (2019). End-to-end differentiable learning of protein structure. *Cell Syst.* **8**:292–301. DOI:10.1016/j.cels.2019.03.006

99. Lee A. C.-L., Harris J. L., Khanna K. K. et al. (2019). A comprehensive review on current advances in peptide drug development and design. *Int. J. Mol. Sci.* **20**:2383. DOI:10.3390/ijms20102383

100. Fosgerau K. and Hoffmann T. (2015). Peptide therapeutics: current status and future directions. *Drug Discov. Today* **20**:122–128. DOI:10.1016/j.drudis.2014.10.003

101. Muttenthaler M. et al. (2021). Trends in peptide drug discovery. *Nat. Rev. Drug Discov.* **20**:309–325. DOI:10.1038/s41573-020-00135-8

102. Agoni C. et al. (2025). Molecular Modelling in Bioactive Peptide Discovery and Characterisation. *Biomolecules* **15**:524. DOI:10.3390/biom15040524

103. Wang J. et al. (2019). Improved modeling of peptide–protein binding through global docking and accelerated molecular dynamics simulations. *Front. Mol. Biosci.* **6**:112. DOI:10.3389/fmolb.2019.00112

104. Dutkiewicz Z. (2022). Computational methods for calculation of protein-ligand binding affinities in structure-based drug design. *Phys. Sci. Rev.* **7**:933–968. DOI:10.1515/psr-2020-0034

105. Seo S. et al. (2024). Pseq2Sites: enhancing protein sequence-based ligand binding-site prediction accuracy via the deep convolutional network and attention mechanism. *Eng. Appl. Artif. Intell.* **127**:107257. DOI:10.1016/j.engappai.2023.107257

106. Ciemny M. et al. (2018). Protein–peptide docking: opportunities and challenges. *Drug Discov. Today* **23**:1530–1537. DOI:10.1016/j.drudis.2018.05.006

107. Kilburg D. and Gallicchio E. (2016). Recent advances in computational models for the study of protein–peptide interactions. *Adv. Protein Chem. Struct. Biol.* **105**:27–57. DOI:10.1016/bs.apcsb.2016.06.002

108. Lee H. et al. (2015). GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization. *Nucleic Acids Res.* **43**:W431–W435. DOI:10.1093/nar/gkv495

109. Xue L. C. et al. (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett.* **589**:3516–3526. DOI:10.1016/j.febslet.2015.10.003

110. Rao R. et al. (2019). Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.* **32**.

111. Zhao Z., Peng Z. and Yang J. (2018). Improving sequence-based prediction of protein–peptide binding residues by introducing intrinsic disorder and a consensus method. *J. Chem. Inf. Model.* **58**:1459–1468. DOI:10.1021/acs.jcim.8b00019

112. Vakser I. A. (2014). Protein-protein docking: From interaction to interactome. *Biophys. J.* **107**:1785–1793. DOI:10.1016/j.bpj.2014.08.033

113. Moal I. H., Moretti R., Baker D. et al. (2013). Scoring functions for protein–protein interactions. *Curr. Opin. Struct. Biol.* **23**:862–867. DOI:10.1016/j.sbi.2013.06.017

114. Hashemifar S., Neyshabur B., Khan A. A. et al. (2018). Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **34**:i802–i810. DOI:10.1093/bioinformatics/bty573

115. Yaffe M. B. (2002). Phosphotyrosine-binding domains in signal transduction. *Nat. Rev. Mol. Cell Biol.* **3**:177–186. DOI:10.1038/nrm759

116. Tinti M. et al. (2013). The SH2 domain interaction landscape. *Cell Rep.* **3**:1293–1305. DOI:10.1016/j.celrep.2013.03.001

117. Chen D. S. and Mellman I. (2013). Oncology meets immunology: the cancer-immunity cycle. *Immunity* **39**:1–10. DOI:10.1016/j.immuni.2013.07.012

118. Mellman I., Coukos G. and Dranoff G. (2011). Cancer immunotherapy comes of age. *Nature* **480**:480–489. DOI:10.1038/nature10673

119. Dunn G. P., Old L. J. and Schreiber R. D. (2004). The three Es of cancer immunoediting. *Annu. Rev. Immunol.* **22**:329–360. DOI:10.1146/annurev.immunol.22.012703.104803

120. Schumacher T. N. and Schreiber R. D. (2015). Neoantigens in cancer immunotherapy. *Science* **348**:69–74. DOI:10.1126/science.aaa4971

121. Gomase V. S., Sharma R. and Dhamane S. P. (2025). Innovative immunoinformatics tools for enhancing MHC class I epitope prediction. *Protein Pept. Lett.*. DOI:10.2174/0109298665373152250625054723

122. Liu Y., Ouyang X.-H., Xiao Z.-X. et al. (2020). A review on the methods of peptide–MHC binding prediction. *Curr. Bioinform.* **15**:878–888. DOI:10.2174/1574893615999200429122801

123. Nielsen M. et al. (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein. *PLoS ONE* **2**:e796. DOI:10.1371/journal.pone.0000796

124. Feltkamp M. C. W. et al. (1994). Efficient MHC class I-peptide binding is required but does not ensure MHC class I-restricted immunogenicity. *Mol. Immunol.* **31**:1391–1401. DOI:10.1016/0161-5890(94)90155-4

125. Murugan A. et al. (2012). Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc. Natl. Acad. Sci. U.S.A.* **109**:16161–16166. DOI:10.1073/pnas.1212755109

126. Dash P. et al. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature* **547**:89–93. DOI:10.1038/nature22383

127. Long X. et al. (2025). THLANet: a deep learning framework for predicting TCR–pHLA binding in immunotherapy applications. *PLoS Comput. Biol.* **21**:e1013050. DOI:10.1371/journal.pcbi.1013050

128. Atchley W. R. et al. (2005). Solving the protein sequence metric problem. *Proc. Natl. Acad. Sci. U.S.A.* **102**:6395–6400. DOI:10.1073/pnas.0408677102

129. Pettmann J. et al. (2021). The discriminatory power of the T cell receptor. *eLife* **10**:e67092. DOI:10.7554/eLife.67092

130. Cole D. K. et al. (2014). T-cell Receptor (TCR)-Peptide Specificity Overrides Affinity-enhancing TCR-Major Histocompatibility Complex Interactions. *J. Biol. Chem.* **289**:628–638. DOI:10.1074/jbc.M113.522110

131. Cuendet M. A., Zoete V. and Michielin O. (2011). How T cell receptors interact with peptide-MHCs: A multiple steered molecular dynamics study. *Proteins* **79**:3007–3024. DOI:10.1002/prot.23104

132. Liu Y. C. et al. (2013). Highly divergent T-cell receptor binding modes underlie specific recognition of a bulged viral peptide bound to a human leukocyte antigen class I molecule. *J. Biol. Chem.* **288**:15442–15454. DOI:10.1074/jbc.M112.447185

133. Albert B. A. et al. (2023). Deep Neural Networks Predict MHC-I Epitope Presentation and Transfer Learn Neoepitope Immunogenicity. *bioRxiv*. DOI:10.1101/2022.08.29.505690

134. Weber A., Born J. and Martínez M. R. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* **37**:i237–i244. DOI:10.1093/bioinformatics/btab294

135. Hudson D., Fernandes R. A., Basham M. et al. (2023). Can we predict T cell specificity with digital biology and machine learning? *Nat. Rev. Immunol.* **23**:511–521. DOI:10.1038/s41577-023-00835-3

136. Jurtz V. I. et al. (2018). NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. *bioRxiv*. DOI:10.1101/433706

137. Altman J. D. et al. (1996). Phenotypic analysis of antigen-specific T lymphocytes. *Science* **274**:94–96. DOI:10.1126/science.274.5284.94

138. Zhang S.-Q. et al. (2018). High-throughput determination of the antigen specificities of T cell receptors in single cells. *Nat. Biotechnol.* **36**:1156–1159. DOI:10.1038/nbt.4282

139. Zhang Z. et al. (2021). Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics. *Nat. Methods* **18**:92–99. DOI:10.1038/s41592-020-01020-3

140. Tickotsky N. et al. (2017). McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* **33**:2924–2929. DOI:10.1093/bioinformatics/btx286

141. Yamasaki S. et al. (2006). Mechanistic basis of pre-T cell receptor-mediated autonomous signaling critical for thymocyte development. *Nat. Immunol.* **7**:67. DOI:10.1038/ni1290

142. Reck M. et al. (2016). Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer. *N. Engl. J. Med.* **375**:1823–1833. DOI:10.1056/NEJMoa1606774

143. Rizvi N. A. et al. (2015). Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science* **348**:124–128. DOI:10.1126/science.aaa1348

144. Miao D. et al. (2018). Genomic correlates of response to immune checkpoint therapies in clear cell renal cell carcinoma. *Science* **359**:801–806. DOI:10.1126/science.aan5951

145. Yost K. E. et al. (2019). Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* **25**:1251–1259. DOI:10.1038/s41591-019-0522-3

146. Lin Z. et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**:1123–1130. DOI:10.1126/science.ade2574

147. Elnaggar A. et al. (2023). Ankh: optimized protein language model unlocks general-purpose modelling. *arXiv*:2301.06568.

148. Zhang S., Liu Y. and Xie L. (2022). Physics-aware graph neural network for accurate RNA 3D structure prediction. *arXiv*:2210.16392.

149. Eraslan G., Avsec ., Gagneur J. et al. (2019). Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* **20**:389–403. DOI:10.1038/s41576-019-0122-6

150. Boehr D. D., Nussinov R. and Wright P. E. (2009). The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**:789–796. DOI:10.1038/nchembio.232

151. Changeux J.-P. and Edelstein S. (2011). Conformational selection or induced fit? 50 years of debate resolved. *F1000 Biol. Rep.* **3**:19. DOI:10.3410/B3-19

152. Cocco S. et al. (2018). Inverse statistical physics of protein sequences: a key issues review. *Rep. Prog. Phys.* **81**:032601. DOI:10.1088/1361-6633/aa9965

153. Marks D. S., Hopf T. A. and Sander C. (2012). Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30**:1072–1080. DOI:10.1038/nbt.2419

154. Duan G. and Walther D. (2015). The roles of post-translational modifications in the context of protein interaction networks. *PLoS Comput. Biol.* **11**:e1004049. DOI:10.1371/journal.pcbi.1004049

155. Wang J. et al. (2019). Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* **5**:755–767. DOI:10.1021/acscentsci.8b00913

156. Gainza P. et al. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**:184–192. DOI:10.1038/s41592-019-0666-6

157. Preuer K. et al. (2019). Interpretable deep learning in drug discovery. In: Samek W., Montavon G., Vedaldi A. et al. (eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Springer), pp. 331–345. DOI:10.1007/978-3-030-28954-6_18

158. Mardt A., Pasquali L., Wu H. et al. (2018). VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **9**:5. DOI:10.1038/s41467-017-02388-1

159. Eden E. et al. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**:48. DOI:10.1186/1471-2105-10-48

160. Noid W. G. (2013). Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.* **139**:090901. DOI:10.1063/1.4818908

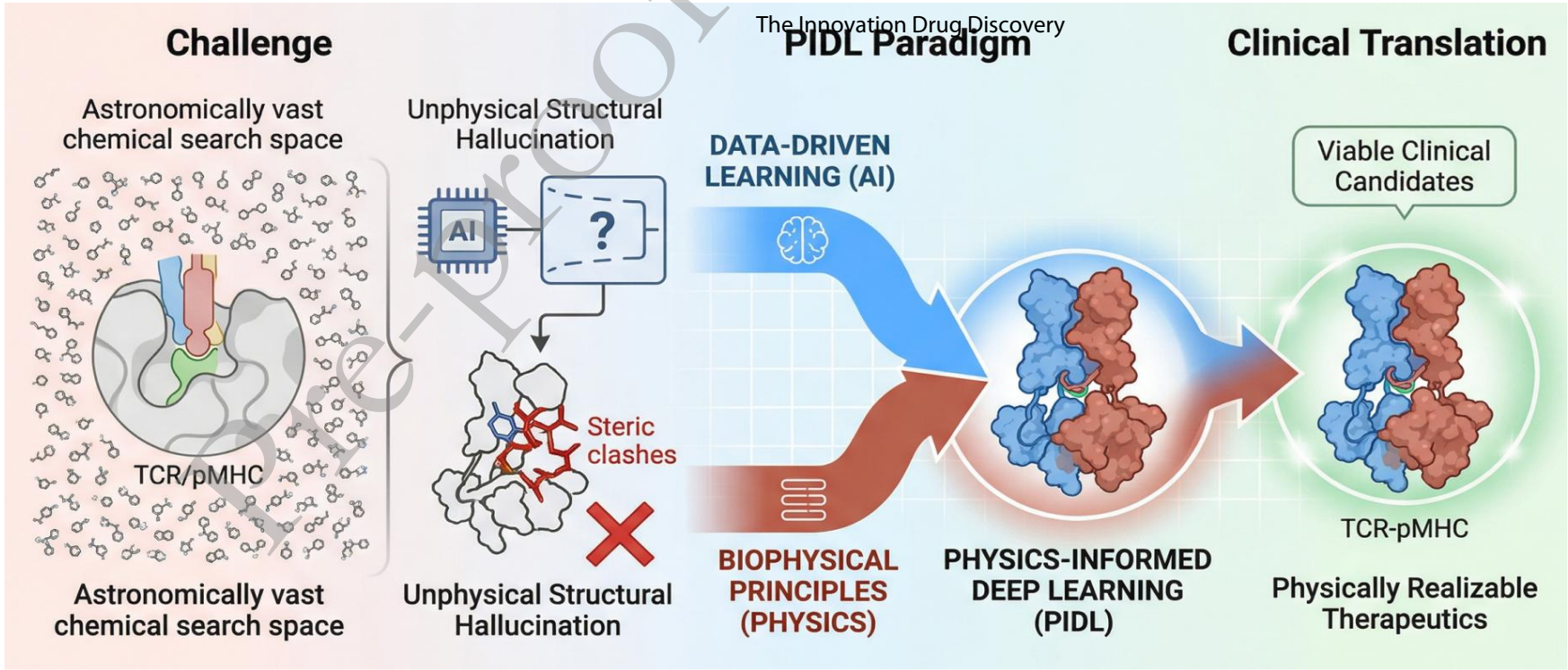
161. Zhang L. et al. (2018). Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**:143001. DOI:10.1103/PhysRevLett.120.143001

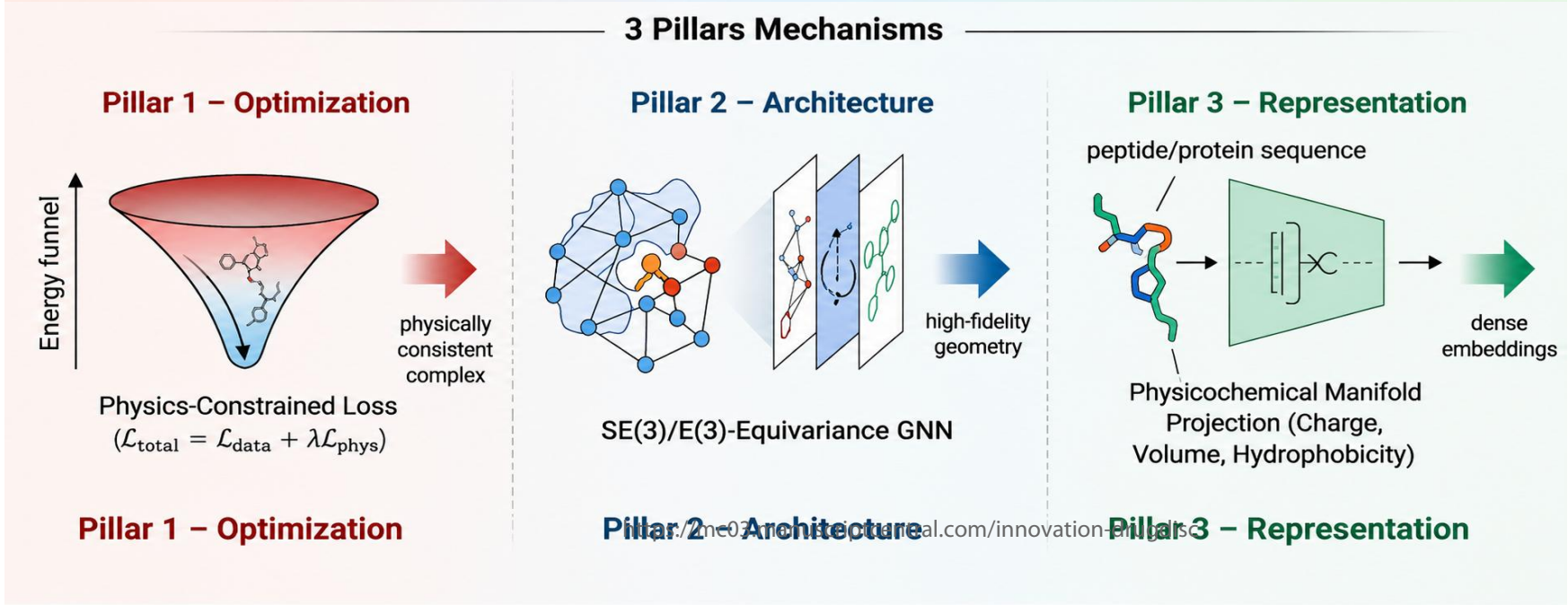
162. Doerr S. et al. (2021). TorchMD: A deep learning framework for molecular simulations. *J. Chem. Theory Comput.* **17**:2355–2363. DOI:10.1021/acs.jctc.0c01343

163. Liu A. et al. (2024). DeepSeek-V3 Technical Report. *arXiv*:2412.19437.

164. Gazizov A. et al. (2026). AF2BIND: predicting small-molecule binding sites using the pair representation of AlphaFold2. *Nat. Methods*:1–10. DOI:10.1038/s41592-026-03011-2

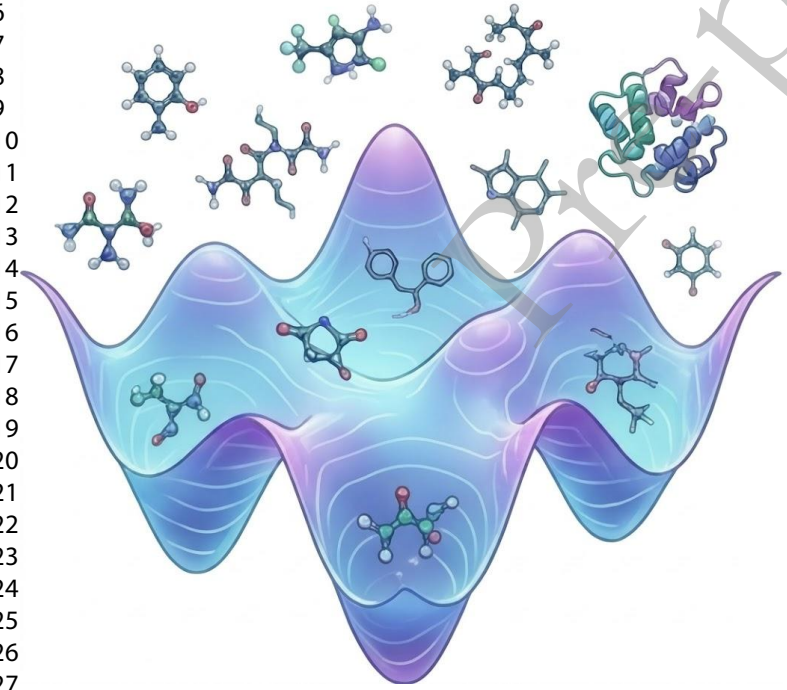
165. Nielsen J. C., Hjorringgaard C., Nygaard M. M. et al. (2024). Machine-learning-guided peptide drug discovery: development of GLP-1 receptor agonists with improved drug properties. *J. Med. Chem.* **67**:11814–11826. DOI:10.1021/acs.jmedchem.4c00417


Challenge
PIDL Paradigm
Clinical Translation
Astronomically vast chemical search space
Unphysical Structural Hallucination
AI
?
DATA-DRIVEN LEARNING (AI)
Viable Clinical Candidates
TCR/pMHC
Steric clashes
TCR-pMHC
BIOPHYSICAL PRINCIPLES (PHYSICS)
PHYSICS-INFORMED DEEP LEARNING (PIDL)
Physically Realizable Therapeutics
Astronomically vast chemical search space
Unphysical Structural Hallucination


3 Pillars Mechanisms
Pillar 1 – Optimization
Energy funnel
physically consistent complex
Physics-Constrained Loss
$(\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{data}} + \lambda\mathcal{L}_{\text{phys}})$
Pillar 2 – Architecture
high-fidelity geometry
SE(3)/E(3)-Equivariance GNN
Pillar 3 – Representation
peptide/protein sequence
dense embeddings
Physicochemical Manifold Projection (Charge, Volume, Hydrophobicity)
Pillar 1 – Optimization
Pillar 2 – Architecture
Pillar 3 – Representation

Background:

Rational Drug Discovery Context



Structural Dark Matter:TCR repertoire diversity exceeds 10^{15}

Modeling biomolecular interactions across vast chemical & structural spaces.

Critical Crossroads:

Two Main Bottlenecks

Physics-Driven Limitations

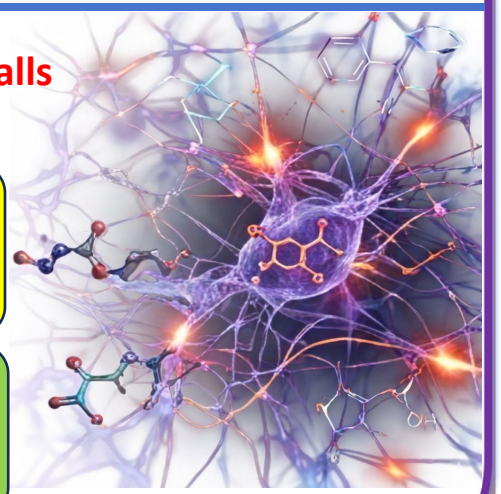
High computational costs & poor scalability

Struggle with dynamic/disordered systems

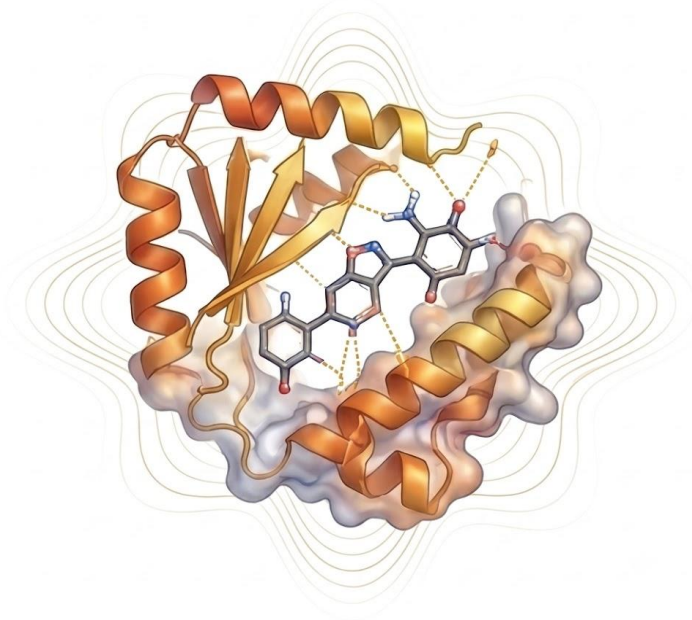
Pure Deep Learning Pitfalls

Black-box nature & lacks physical consistency

Cause severe steric clashes & unstable hallucinations



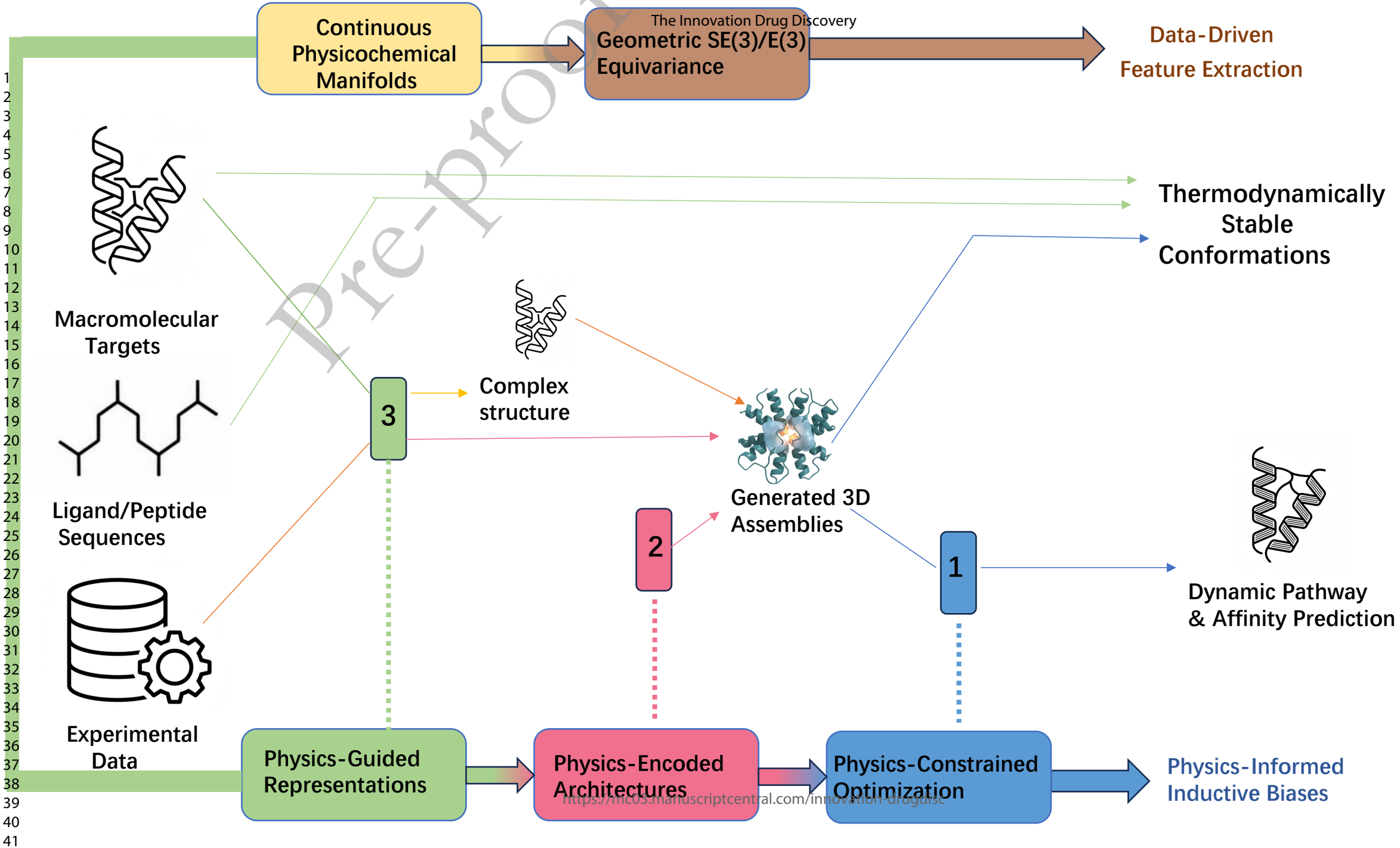
The Paradigm Shift:Physics-Informed AI



☑Embeds physical laws as powerfull inductive biases.

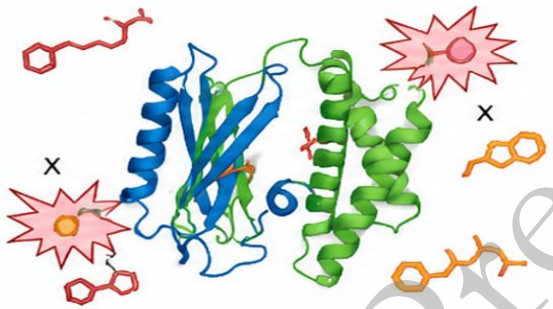
☑Guarantees physically realizable & thermodynamically stable structures.

☑Unifies physical priors with data-driven learning.



Inputs	Core Paradigms & Mathematical Mechanisms	Outputs

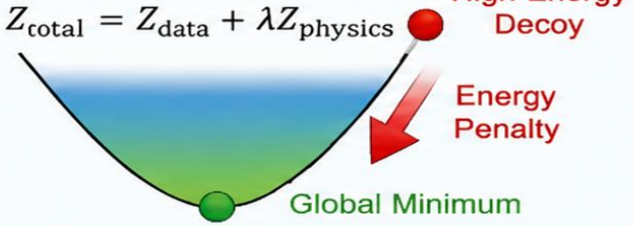
Generated Candidates & Decoys



Energy-Based Models (EBMs)

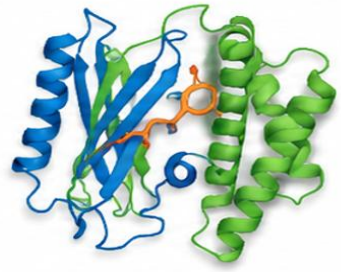
$$P(\mathbf{x}) = \frac{\exp(-\beta E_\theta(\mathbf{x}))}{Z(\theta)}$$

$$Z_{\text{total}} = Z_{\text{data}} + \lambda Z_{\text{physics}}$$



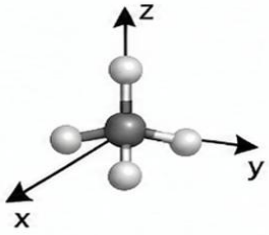
Docking/ Sampling

Thermodynamically Stable Conformations



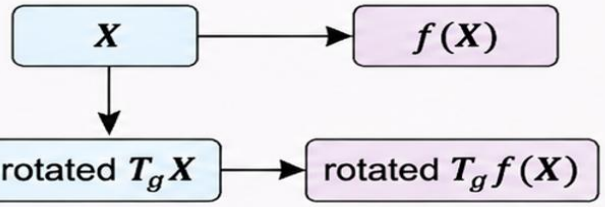
Physics-Encoded Architectures

3D Molecular Graphs (X)



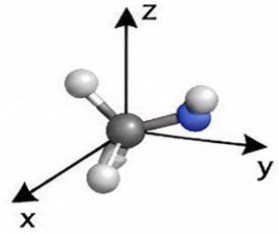
SE(3)/E(3) Equivariant Message Passing

$$f(T_g X) = T_g f(X),\ \forall T_g \in SE(3)$$



Rotational Transformation

Equivariant Features & Rotated Coordinates

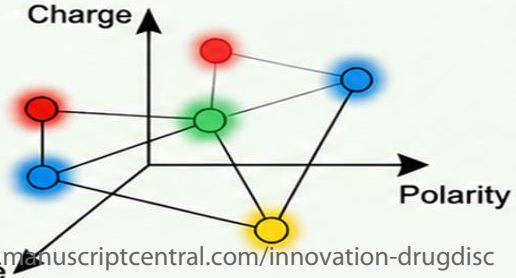


Physics-Guided Representations

Discrete Biological Sequences (S)

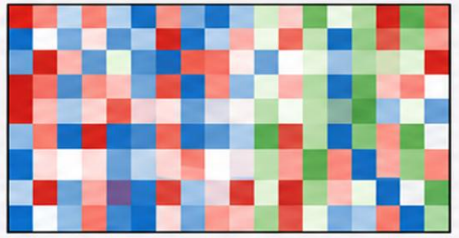
Physicochemical Manifold Projection

$$M: S \rightarrow \text{Physical Priors} \rightarrow \mathbb{R}^{d_{phys}}$$



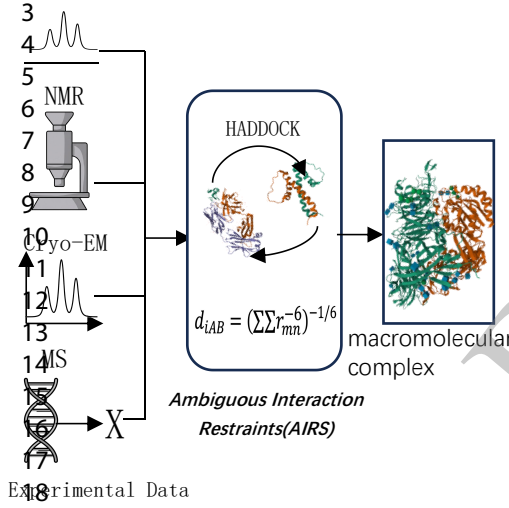
Projection/ Embedding

Continuous Physicochemical Embeddings

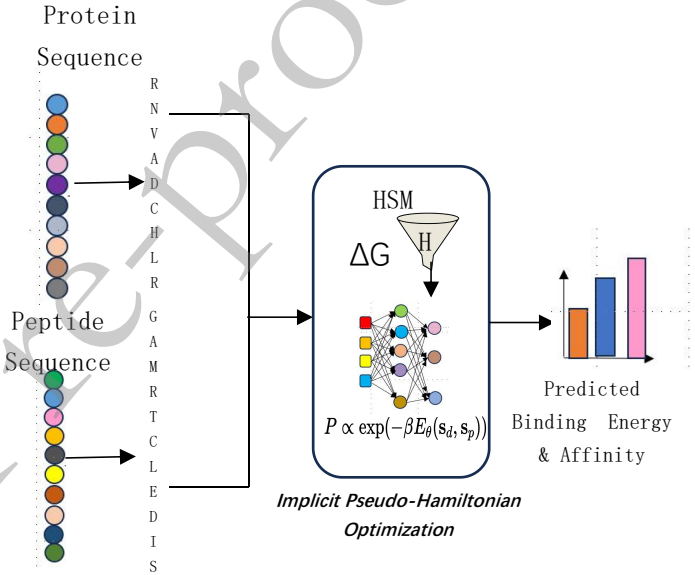


A Physics-Constrained Docking

Integrative Modeling via Explicit Restraints(HADDOCK)

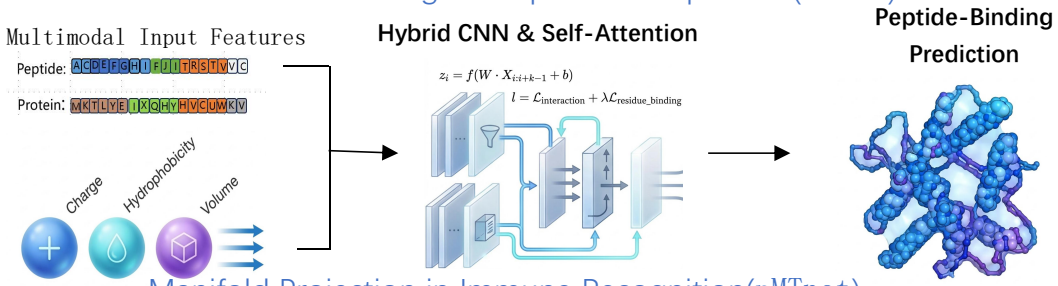


Statistical Mechanics via EBMs(HSM)

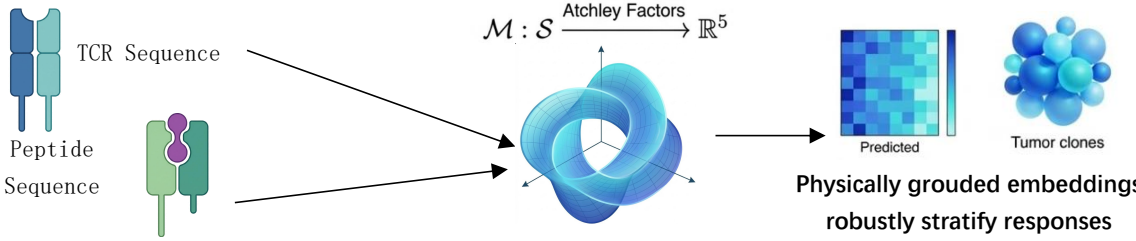


B Physics-Guided Representations

Multi-Task Learning for Peptide Therapeutics (CAMP)



Manifold Projection in Immune Recognition(pMTnet)

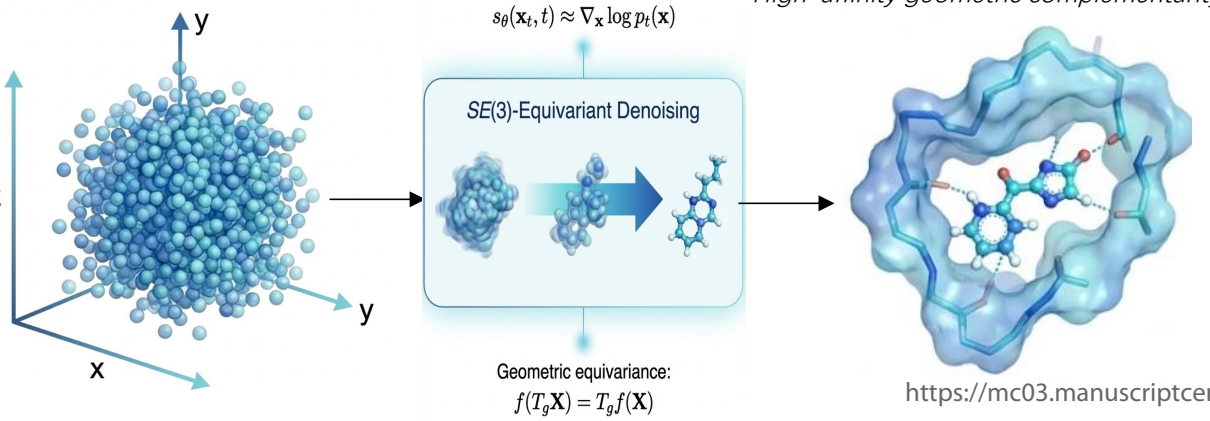


C Physics-Encoded Architectures

Equivariant Diffusion Model(eg. DiffDock,AlphaFold 3, RoseTTAFold All-Atom (RFAA) & Chai-1)

Random Coordinate Cloud

High-entropy noise state



Thermodynamically Stable Pose

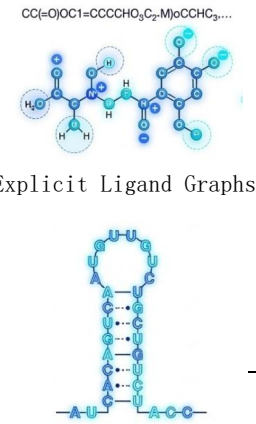
High-affinity geometric complementarity

SE(3)-Equivariant Denoising

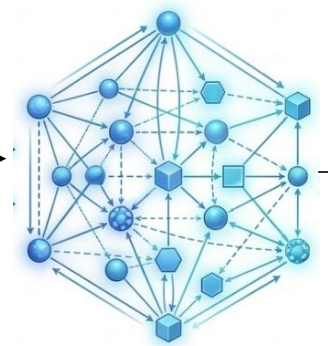
D Universal All-Atom Assemblies

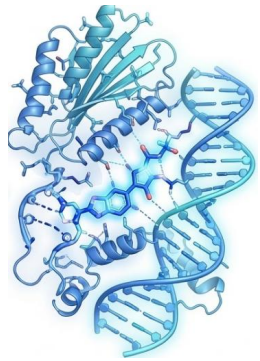
Distinct Physical Priors

All-Atom Complex Prediction



Unified Heterogeneous Graph





Geometic Base-Pairing Topologies

Heterogeneous Graph Neural Network (Heterogeneous GNN)

Processing diverse molecular alphabets natively

All-Atom Complex Prediction

Bridging small-molecule docking and nucleic acid folding

Evolutionary Crossing

Ontological Restructuring:*From Static to Dynamic Systems*

Scale & Environment

Crystallographic Bias & In Vacuo Models

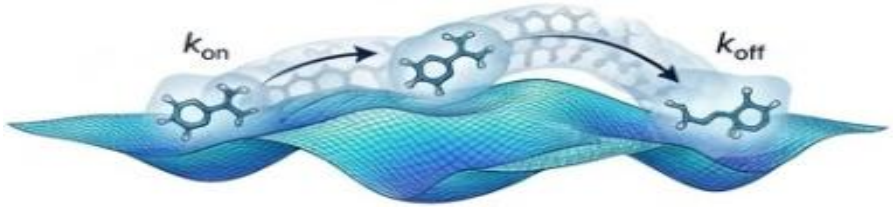
Context Integration

Context-Aware Mesoscale Interactomes

Thermodynamics & Time

Entropic Blindness & AI Hallucinations

Dynamic Simulation

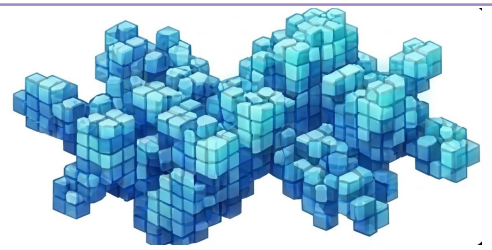


4D Spatiotemporal Pathway Generation

Algorithmic Architecture

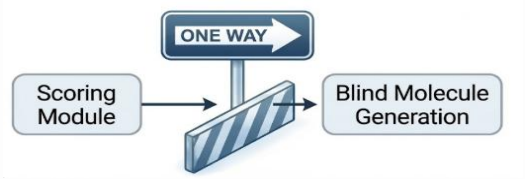
Sequence Truncation & Complexity Bottleneck

Visual Token Compression



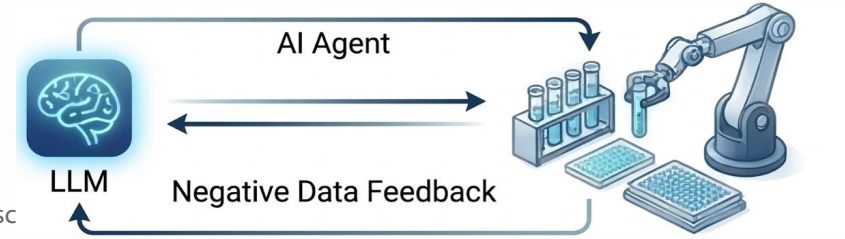
Vision-Inspired Full-Scale Encoding

Discovery Loop



Discriminative/Generative

Autonomous Feedback



Lab-in-the-Loop Autonomous Discovery

LICENSE TO PUBLISH

The corresponding author listed as the "Corresponding Author" on the submission site (the "Author") is executing this license to publish (the "License") to Innovation Press with its principal place of business at Room 1006,10/F., Po Yip Building, 23 Hing Yip Street, Kwun Tong, Kowloon, Hong Kong (the "Publisher"), and each of them, along with the Author, constitutes a "Party" (the "Parties").

1. License Granting

This License gives the author the sole, worldwide, royalty-free, perpetual (if the relevant copyright is in effect) right and license to use the Work (as specified in submission system) for any commercial or educational purpose, including but not restricted to publishing, duplicating, marketing, distributing (both directly and through distributors), subleasing, and selling copies of the Work internationally for the Term. Note: For government employees in different countries, a statement declaring that the viewpoints stated do not necessarily represent those of the military or government agency for which they work may be required for work performed by government employees or military personnel. Please make sure to consult the relevant authorities and incorporate the required statement into the body of your contribution.

2. Warranties, Indemnification, and Limitation of Liability

a. The Authors warrant and represent to the publisher that:

1. The Author possesses the authority and capability to execute this License, to bestow the rights and licenses under it, and to fulfill all the additional duties outlined in it;
2. The person executing this License on the Author's behalf is allowed to do so, and the Author has not previously assigned, transferred, or otherwise encumbered the rights or licenses granted pursuant to this License;
3. The Work and the licenses granted herein do not and will not violate, misappropriate, or infringe upon any proprietary rights, intellectual property rights, contracts, or other third-party rights or interests;
4. In the event that the work is co-authored, the author has received written consent from each co-author to accept this license on their behalf, and each co-author has read, comprehended, and accepted its conditions;
5. All necessary releases and permissions to incorporate works and materials into the Work, including those from patients whose names or likenesses are submitted, have been obtained by the Author; all such releases and permissions are in full force and effect, and the Author will promptly furnish the Publisher with any such release or permission upon request from the Publisher;
6. Except for the Journal (as defined in Section 3), neither the Work nor any of the content it contains, in whole or in part, has been published or is being considered for publication;
7. Prior to or concurrently with the submission of the Work, the Author has given the Publisher notice of all intellectual contributions, technical assistance, financial or material support, and any other relationships that could create or result in a conflict of interest;
8. The copyright of the Author and each additional author of the Work is the only copyright that applies to the Work;
9. The Work does not and will not infringe upon the rights of any third party to privacy or publicity, or defame

or disparage any third party;
10. The Work does not and will not include any content that is scandalous, indecent, or carelessly created;
11. The Work has not been and will not be falsely claimed, copied, or misattributed;
12. There is no element of the Author's personal or professional circumstances that might give rise to a conflict of interest with regard to the Work, either now or in the previous 12 months;
13. Nothing of value has been received by the Author, any member of the Author's immediate family, or any person or organization with which the Author has or has had a substantial working relationship from a commercial party that is connected to the Work's subject matter directly or indirectly.
b. The Author agrees to defend and hold harmless the Publisher, its directors, officers, employees, agents, and representatives from and against any and all liability, damage, loss, costs, or expenses (including reasonable attorney's fees and settlement costs) incurred by any such party arising out of, or relating to, any misrepresentation in, or breach or alleged breach of the Author's representations or warranties in this License. The Author does this with this agreement. The parties that are indemnified, or any of them, may take up the defense at the Author's expense if the Author neglects to act in a timely or diligent manner in defending any of the parties. Should this License terminate or expire, the duties under this indemnity shall continue.

3. Creative Commons Licenses

The Author agrees and acknowledges that the Work will be published in Innovation (the "Journal") by the Publisher and made freely accessible to users under the terms of one of the following Creative Commons Licenses; the Author will choose which license to apply in the submission system:
(i) The current version of the Creative Commons Attribution 4.0 International Public License (the "CC BY") may be seen at http://creativecommons.org/licenses/by/4.0/legalcode. The Publisher is the only “Licensor” of the Work, as that term is defined in the CC BY, and the Author acknowledges and accepts that the Publisher may make the Work freely available to all users in accordance with the CC BY.

4. Royalties

The Author understands and accepts that there are no royalties or other costs associated with this License. The author releases the publisher and its sublicensees from all rights to royalties or other payments related to the work or any use of the work, to the fullest extent allowed by law.

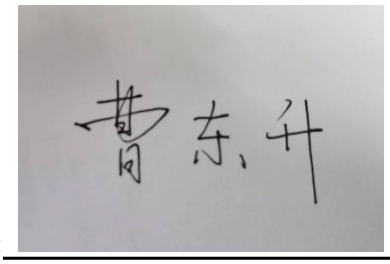
5. Miscellaneous

1. Assignment. The Author may not assign or transfer this License, in whole or in part. This License may be freely assigned by the Publisher. This License will benefit the parties hereto, as well as each of their respective successors and authorized assignee, and be binding upon them.
2. Enforcement. For the purposes of this License, signatures in the Portable Document Format (PDF) or via facsimile will be considered original signatures.
3. Complete Agreement; Modification. Regarding the rights and obligations outlined herein, this License contains the whole understanding between the Parties and supersedes all prior or contemporaneous oral or written representations and agreements. It cannot be changed or updated without the Publisher's express written consent.
4. Governing Law. This License shall be governed in all respects by the laws of Hong Kong SAR, China, without regard to its conflict of laws rules.

5. Headings. All headers are only for reference and have no bearing on how any of the provisions inside are understood or interpreted.
6. Severability. Should a court of competent jurisdiction rule that any term of this License is illegal, void, or unenforceable under current or future laws, that provision will be completely severable and will only be enforced to the extent authorized by existing law. This License shall be interpreted and enforced in all cases as though the illegal, invalid, or unenforceable provision had never been a part of it; the remaining provisions of this License shall be unaffected by the illegal, invalid, or unenforceable provision, or by its severance from this License, and shall remain in full force and effect.
7. The Parties' current state. The Parties operate as separate businesses. No agency, joint venture, partnership, or fiduciary relationship between the Parties is intended to be formed or established by anything in this License, and neither Party has the power to bind the other Party or be held accountable for the actions or inactions of the other Party.
8. Amendment and Waiver. The Publisher's failure to assert a breach of any term of this License or its waiver of any prior breach shall not be construed as a waiver of any subsequent breach, nor shall it in any manner impair the continued efficacy of any such provision. This License cannot be waived except by a written agreement from the Publisher on all terms and conditions.
9. Trial by Jury Waiver. **In connection with any disagreement or legal action arising out of this agreement or the subject matter hereof, the author thus waives the right to a jury trial.**

Manuscript Title: Physics-Informed Deep Learning: A New Paradigm for Macromolecular Recognition and Rational Protein Design

All Authors: Haobo Xie, Hao Wang, Xiaojun Yao, Jinyan Li, Yi Pan, Aiping Lu, Xiangxiang Zeng, Dongsheng Cao* , Xiangzheng Fu*

Signature: 

(**All corresponding authors** sign here using electronic signature on behalf of all authors.)

Date:July 13, 2026