ChemComm

FEATURE ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Chem. Commun.*, 2026, **62**, 752

Computational design of protein complexes: influence of binding affinity

Fathima Ridha, K. Harini, N. R. Siva Shanmugam, Rahul Nikam and M. Michael Gromiha *

The interaction of proteins with diverse molecular partners, including other proteins, nucleic acids, and carbohydrates, is essential for performing various functions, from signal transduction and gene regulation to immune recognition and cellular transport. These interactions are largely governed by the three-dimensional structures and dynamics of biomolecular complexes, which in turn dictate their binding affinities and functional specificity. While recent advances in AI-driven structure prediction have greatly improved our ability to model such complexes, accurately predicting and engineering their binding affinities remains a key challenge. In this article, we review emerging computational strategies for affinity prediction and rational design across protein–protein, protein–DNA/RNA, and protein–carbohydrate complexes. We discuss the role of machine learning and deep learning in advancing structure-based and sequence-based affinity models, assess current databases and benchmarks, and highlight recent tools for predicting the effects of mutations on binding affinity. We conclude by discussing future opportunities at the intersection of AI, high-throughput screening, and data-driven modeling to enable affinity-guided design of functional biomolecular assemblies.

Received 21st August 2025,
Accepted 19th November 2025

DOI: 10.1039/d5cc04821d

rsc.li/chemcomm

1. Introduction

Proteins interact with other macromolecules, including proteins, nucleic acids, carbohydrates, as well as small ligands, and are fundamental to most of the cellular processes.[1,2] These interactions underlie crucial biological phenomena such as signal transduction, transcriptional regulation, immune recognition,

Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600036, India.
E-mail: gromiha@iitm.ac.in

Fathima Ridha

Dr Fathima Ridha earned her PhD in Computational Biology from the Indian Institute of Technology (IIT) Madras, under the supervision of Prof. M. Michael Gromiha. Her doctoral research focused on membrane protein–protein interactions, developing computational methods to predict binding affinity in membrane protein complexes and their mutants and to understand the molecular basis of disease-associated variations. She has begun her postdoctoral research at the Max Planck Institute for Multidisciplinary Sciences in Germany, where she applies statistical and computational biology approaches to study microbial protein function and molecular mechanisms.

K. Harini

Dr K. Harini completed her PhD under the guidance of Prof. M. Michael Gromiha, in the Department of Biotechnology at the Indian Institute of Technology Madras, India. Her research mainly focused on understanding and predicting the binding affinities of protein–nucleic acid complexes. She also investigated the effect of mutations on the binding affinities of the protein–DNA/RNA complexes. She is currently a postdoctoral researcher at the Indian Institute of Technology Bombay, where she is continuing her work by analyzing ChIP sequencing data to advance the understanding of yeast biology.

and enzymatic catalysis.[3] Specific examples of such interactions are found across all domains of biology. For instance, the formation of immune complexes through antibody–antigen recognition is crucial for pathogen defense, while protein–DNA recognition underlies gene regulation by transcription factors. Protein–protein complexes, such as those involved in cell signaling pathways (*e.g.*, kinase–substrate pairs), and protein–carbohydrate interactions that determine cell–cell adhesion or host–pathogen recognition, further highlight the diversity and importance of these interactions in both normal physiology and disease. The specificity, stability, and function of these complexes are largely determined by their three-dimensional structures and, crucially, by their binding affinity, a quantitative measure describing the strength of interaction.[4]

Binding affinity not only underpins our understanding of the molecular recognition and biological function but also plays a decisive role in drug discovery, synthetic biology, and protein engineering.[5,6] Even small changes in binding affinity, often caused by mutations at the interface residues, can lead to profound changes in the cellular functions, disease susceptibility, or therapeutic response.[7,8] Experimentally, binding affinities are measured using methods such as isothermal titration calorimetry (ITC), surface plasmon resonance (SPR), and fluorescence-based assays. However, these techniques are labor-intensive and not feasible for high-throughput or proteome-scale analyses.[8]

N. R. Siva Shanmugam

Dr N. R. Siva Shanmugam is currently a postdoctoral research associate at the University of Nebraska–Lincoln, United States. He earned his PhD in Computational Biology under the guidance of Prof. M. Michael Gromiha in the Department of Biotechnology at the Indian Institute of Technology Madras (IIT-M), India, and received his MTech degree in Bioinformatics from Bharathidasan University, India. His research focuses on the binding affinity of protein–carbohydrate complexes, utilizing sequence and structural features, as well as on predicting carbohydrate-binding proteins.

Rahul Nikam

Dr Rahul Nikam is a Postdoctoral Researcher at the University of Michigan, specialising in computational biology, with a research focus on metagenomic and metatranscriptomic analysis of the oral microbiome. He earned his PhD from the Indian Institute of Technology (IIT) Madras under the mentorship of Prof. M. Michael Gromiha. Dr Nikam has developed innovative computational tools, including Seq2-Feature, DeepBSRPred, and DeepPPAPredMut, and contributed to the establishment of database resources such as ProThermDB. Through the integration of large-scale sequencing data, his ongoing work aims to advance understanding of microbial composition, gene expression, and functional dynamics within the oral ecosystem.

M. Michael Gromiha

Dr M Michael Gromiha is working as a Professor at the Department of Biotechnology, Indian Institute of Technology (IIT) Madras, India. His main research interests include protein structure and function, stability and binding affinity of protein complexes and their mutants, disease-causing mutations, structure based drug design, next generation sequence analysis, and development of bioinformatics databases and AI/ML based tools. He has published more than 280 research articles, 70 reviews, eight editorials, and three books. His papers received more than 17,000 citations and his h-index is 64. He has guided 25 PhD students, 15 Post-docs and handled more than 20 national and international projects. He is an Associate Editor of BMC Bioinformatics, Bioinformatics Advances, Frontiers in Bioinformatics and Bioinformatics and Biology Insights as well as an Editorial board member of Scientific Reports, Biology Direct, Journal of Bioinformatics and Computational Biology, Genes and Current Computer Aided Drug Design. He has received several awards including the Oxford University Press Bioinformatics prize, ICTP Associateship Award, ICMR International Fellowship for Senior Biomedical Scientists, INSA Senior Scientist Award, Institute Research and Development Award from IIT Madras, Outstanding Performance Award, Tokyo Institute of Technology, Japan, Tamilnadu Scientist Award and ASC Masila Vijaya Award for excellence in research and publications. He is ranked as one of the topmost 0.5% of highly cited researchers in the world. He is an elected Fellow of Indian National Science Academy (FNA) and Academy of Sciences, Chennai.

The field of computational affinity prediction has advanced rapidly, propelled especially by the convergence of machine learning, deep learning, and structure modeling. The introduction of AI-powered structure prediction tools such as AlphaFold2/Multimer[9,10] and RoseTTAFoldNA,[11] along with more recent developments like AlphaFold3,[12] enables atomic-resolution modeling of protein–protein, protein–nucleic acid, and protein–carbohydrate complexes. These advances drive computational workflows that extend beyond structure reconstruction to the estimation and optimization of binding affinities.[13] Modern strategies for binding affinity predictions have evolved from classical energy calculation, such as molecular mechanics Poisson–Boltzmann/surface area (MM/PBSA) methods[14] to machine learning/deep learning approaches harnessing sequence, structure, and evolutionary data.[15–18] Current state-of-the-art methods apply graph neural networks, transformer-based language models, and diverse ensembles for affinity prediction across complex types.[19,20]

In parallel, the development of well-curated binding affinity databases such as SKEMPI,[21] PDBbind,[22] PROXiMATE,[23] MPAD,[24] ProNAB,[25] and ProCaff[26] has been pivotal in supporting not only model training and benchmarking but also a deeper understanding of the thermodynamics of molecular recognition. These resources are important for refining energy functions, training machine-learning models, and designing novel protein interfaces.[27,28]

Despite these advances, accurately predicting binding affinities remains challenging due to the complex interplay of enthalpic and entropic contributions, solvent effects, and conformational flexibility at biomolecular interfaces.[29] Traditional methods often struggle to balance computational efficiency with physical accuracy, while machine learning approaches depend heavily on the quality, diversity, and representativeness of available data. Moreover, different types of macromolecular complexes (protein–protein, protein–nucleic acid, and protein–carbohydrate) exhibit distinct physicochemical determinants of binding, further underscoring the difficulty of developing broadly generalizable frameworks for affinity prediction.

Over the past several years, our group has contributed to this field through the development of comprehensive, literature-derived databases of experimentally determined binding affinities for protein–protein,[23,24] protein–nucleic acid,[25] and protein–carbohydrate complexes,[26] which are widely used for benchmarking and training predictive models. In addition, we have developed machine learning and deep learning–based computational tools for predicting both wild-type affinities[30–34] and mutation-induced changes in binding affinity of protein complexes.[7,20,35–37]

In this review, we explore advances in the computational design and prediction of protein complexes with a focus on binding affinity. We describe current strategies for affinity prediction across protein–protein, protein–nucleic acid (DNA/RNA), and protein–carbohydrate complexes, and highlight recent contributions from machine learning and AI. We also survey key databases supporting the field, with special attention to computational methods that predict mutational effects on binding affinity, a crucial need for understanding disease mechanisms and therapeutic engineering. We conclude with a discussion of future directions, including AI-driven structure prediction, high-throughput computational screening, and novel affinity-aware design platforms.

2. Databases for binding affinities of protein complexes

Quantitative data on biomolecular binding affinities are central to understanding the thermodynamics of molecular recognition. Such measurements are typically derived from biophysical experiments, which enable one to compute thermodynamic parameters such as dissociation constant (K_d), binding free energy (ΔG), or change in affinity upon mutation ($\Delta\Delta G$). However, this information is often scattered across publications and supplemental materials, making centralized, curated databases indispensable for accessibility, consistency, and downstream applications.

Over the past two decades, several databases have been developed to catalog experimentally measured binding interactions across a wide range of biological systems, namely, protein–protein, protein–nucleic acid, and protein–carbohydrate complexes. A comparison of major binding affinity databases, including their data types, coverage, and features, is provided in Table 1. These resources have become essential not only for uncovering the physical principles of binding but also for enabling computational modeling, mutational analysis, and protein engineering.

Early efforts were largely focused on protein–protein interactions, with databases like SKEMPI,[21] PDBbind,[22] and PROXiMATE,[23] providing quantitative affinity data alongside structural information. PROXiMATE, developed by our group, is a curated database of thermodynamic effects of missense mutations in heterodimeric protein–protein complexes, enriched with sequence, structural, and functional annotations; it also includes binding affinity data for homodimeric complexes. These databases are critical for training machine learning prediction models, benchmarking scoring functions, and guiding rational protein design. The binding affinity values enable comparative assessment of complex strength.

Recently, protein–nucleic acid interactions have received growing attention, with resources like PDBbind,[22] ProNIT,[43] and dbAMEPNI[44] compiling measured affinities along with detailed annotations. Although smaller in scale compared to protein–protein datasets, these resources are critical for understanding sequence- and structure-specific recognition of DNA and RNA targets by proteins. To address limitations in coverage and consistency, we developed ProNAB,[25] currently the largest database of experimentally measured affinities for wild-type and mutant protein–DNA/RNA complexes. ProNAB enables more systematic analysis of binding energetics and supports the development of predictive models tailored to nucleic acid recognition.[48]

Protein–carbohydrate interactions, despite their biological significance, have remained underrepresented in quantitative

Table 1 Databases for binding affinities of protein complexes

Database	Interaction type	Data available	URL
SKEMPI v2.0[21]	Protein–protein	Wild-type and mutant ΔG for structure-known complexes	https://life.bsc.es/pid/skempi2/
PROXiMATE[23]	Protein–protein	Binding affinity (K_d/ΔG/$\Delta\Delta G$) for complexes with both known and unknown structures	https://www.iitm.ac.in/bioinfo/PROXiMATE/
Affinity Benchmark v5.5[38]	Protein–protein	Benchmark set of protein–protein complexes with experimentally measured K_d and corresponding calculated ΔG values.	https://bmm.cancerresearchuk.org/~bmmadmin/Affinity[a]
AB-Bind[39]	Protein–protein	Affinity changes upon mutation ($\Delta\Delta G$) in antibody–antigen complexes	https://github.com/sarahsirin/AB-Bind-Database/tree/master
SAbDab[40]	Protein–protein	Structural database of antibody–antigen complexes; a few with affinity data	https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/sabdab
ATLAS[41]	Protein–protein	T cell receptor–peptide–MHC (TCR–pMHC) complexes, linking experimentally determined binding affinities with structural information.	https://zlab.umassmed.edu/atlas/web/
PPB-Affinity[42]	Protein–protein	Experimental affinities for protein–protein complexes compiled from SKEMPI v2.0, AB-Bind, SAbDab, PDBbind v2020, Affinity Benchmark v5.5, ATLAS	https://github.com/Huatsing-Lau/PPB-Affinity-DataPrepWorkflow
MPAD[24]	Protein–protein	Binding affinity (K_d/ΔG/$\Delta\Delta G$) specific for membrane protein–protein complexes, along with membrane-based features	https://web.iitm.ac.in/bioinfo2/mpad/
ProNIT[43]	Protein–nucleic acid	Experimentally determined thermodynamic data (K_d, ΔG, ΔH, ΔS)	https://www.rtc.riken.go.jp/jouhou/pronit/pronit.html[a]
dbAMEPNI[44]	Protein–nucleic acid	Alanine mutations ($\Delta\Delta G$)	https://zhulab.ahu.edu.cn/dbAMEPNI[a]
ProNAB[25]	Protein–nucleic acid	Dissociation constant (K_d), free energy change (ΔG), and its change upon mutation ($\Delta\Delta G$), along with experimental conditions	https://web.iitm.ac.in/bioinfo2/pronab/
PNATDB[45]	Protein–nucleic acid	Includes molecular interactions information along with binding affinity	https://chemyang.ccnu.edu.cn/ccb/database/PNAT/
ProCaff[26]	Protein–carbohydrate	Dissociation constant (K_d), Gibbs free energy (ΔG, $\Delta\Delta G$), experimental conditions, sequence, structure, and literature information	https://web.iitm.ac.in/bioinfo2/procaff/
CarbDisMut[46]	Protein–carbohydrate	Disease-causing mutations in human carbohydrate-binding proteins and predicted free energy change upon mutations ($\Delta\Delta G$) using PCA-MutPred	https://web.iitm.ac.in/bioinfo2/carbdismut/
ProCarbDB[47]	Protein–carbohydrate	Structural database of protein–carbohydrate complexes, some with affinity/mutation data	https://www.procarbdb.science/procarb/
PDBbind+[22]	Protein–protein, protein–nucleic acid, protein–ligand, and nucleic acid–ligand	Binding affinity data for biomolecular complexes in PDB	https://www.pdbbind-plus.org.cn/

[a] Not accessible; last accessed on 05 August 2025.

binding datasets. To address this gap, we developed ProCaff,[26] the first curated database of experimental binding affinity of protein–carbohydrate complexes and their mutants, collected from the literature. ProCaff is a valuable resource to gain insights for understanding the importance of specific interactions at the interface of protein–carbohydrate complexes and the recognition mechanism of protein–carbohydrate complexes.

Membrane protein–protein interactions, despite their central role in signaling and therapeutic targeting, remain underrepresented in affinity databases, with data scattered across the literature and no comprehensive resource available until recently. We addressed this with MPAD,[24] the first dedicated database of binding affinities for membrane protein complexes and their mutants, featuring over 5400 curated entries along with membrane-specific features. This resource enables the exploration of energetics in membrane protein complexes and the impact of mutations on binding affinity, providing deeper insights into disease mechanisms and supporting the development of more effective, targeted therapies.

To ensure consistency and reliability across our datasets, we adopted a unified curation pipeline across all our databases, including PROXiMATE, ProNAB, ProCaff, and MPAD. Relevant research articles were identified from PubMed and major journal websites using combinations of keywords related to target proteins, structural and functional classes, binding energetics (*e.g.*, K_d, ΔG, $\Delta\Delta G$), and experimental techniques (*e.g.*, ITC, SPR). Each article was manually reviewed to extract binding affinity values, experimental conditions, structural details (PDB ID), and literature metadata. We also integrated data from existing databases (*e.g.*, PDBbind, SKEMPI, *etc.*) by mapping protein IDs and, wherever available, enriching incomplete records with missing details such as experimental methods, data location, *etc.*, and standardized all thermodynamic parameters to ensure comparability. Additional context-specific annotations, such as membrane-specific features for MPAD, were included to enrich downstream analysis. Finally, each dataset was organized into a user-friendly database featuring robust search, filter, and visualization options, along with

provisions for data upload and download, enabling easy access and utility for the broader research community.

Altogether, these databases represent significant progress toward making quantitative affinity data accessible across diverse classes of biomolecular complexes. As experimental techniques and computational methods continue to evolve, regular updates and broader coverage will be critical. Looking ahead, an integrative platform combining thermodynamic, structural, and contextual annotations across diverse interaction types would substantially enhance accessibility and interoperability. Mapping all known interaction data for a given protein could offer a comprehensive view of its recognition landscape and support the development of more generalizable predictive models.

3. Computational algorithms for predicting the binding affinity

Although experimental methods are crucial for accurately validating biomolecular binding affinities, they are often impractical for large-scale screening due to their high demands on time and resources. Consequently, the available data on the binding affinity could be efficiently leveraged for the development of computational methods to screen, identify, and prioritize potentially novel protein–protein/nucleic acid or carbohydrate interactions. The emergence of such *in silico* models can significantly speed up the screening process and reduce the resources and cost spent on eliminating false positives. Precise prediction of binding affinities is also fundamental to engineering proteins with high specificity and strong affinity toward their targets. A variety of computational approaches have been developed for binding affinity prediction across protein–protein, protein–DNA/RNA, and protein–carbohydrate complexes.

Protein–protein binding affinities are commonly predicted by leveraging a combination of sequence information and structure-based features. These predictions are achieved using a range of computational approaches, including traditional machine learning algorithms to more recent deep learning frameworks. Sequence-based methods typically utilize fundamental physicochemical amino acid properties, sequence conservation, and residue-level probability estimates for binding site involvement. Yugandhar and Gromiha[15] developed PPA-Pred, a sequence-based prediction method, and demonstrated that stratifying datasets according to protein functional classes led to a notable improvement in predictive accuracy.

Structure-based methods have been developed to incorporate geometric and interaction-specific features. These include features such as surface area[49] and interfacial contacts,[50] which provide detailed insights into the physical nature of protein–protein interfaces. With the advent of deep learning algorithms, the performance of the models is significantly improved with the incorporation of graph neural networks, and protein language models, trained on large-scale sequence data, introduce contextual features from raw protein sequences.[13,51] Moreover, advancements in structure prediction tools such as AlphaFold-multimer, revolutionized the field by enabling the high-confidence prediction of three-dimensional protein complexes directly from sequence data. These predicted structures can be used to derive structural features, thereby enhancing model robustness and predictive power.[30] Most recently, Ridha and Gromiha[31] developed a prediction model specific to membrane proteins, utilizing the structural and sequence-based features to address the unique challenges posed by these complex structures. Table 2 lists different statistical and machine learning models for predicting binding affinities along with the sequence/structure-based features employed for prediction.

The prediction of binding affinities in protein–nucleic acid complexes is also foundational for understanding gene regulation. These prediction methods employ a range of computational approaches, including molecular dynamics simulations, statistical methods, and machine learning techniques.[64] Similar to protein–protein interaction studies, structure-based parameters such as buried surface area and interatomic contacts have been identified as key determinants of binding affinity of protein–nucleic acid complexes. In addition, energetic parameters such as contact potentials and electrostatic interactions are related to the binding affinities.[32] Further, on the nucleic acid side, features such as base-step parameters, secondary structures, and local structural motifs are known to be important for nucleic acid recognition.[33] As a concrete example, Pant *et al.*[64] demonstrated that bicyclo-nucleotide modifications in DNA increase the affinity of protein–DNA complexes. Interestingly, intrinsically disordered regions have been found to significantly influence the binding strength of DNA-interacting proteins.[65] Barissi *et al.*[56] developed a random forest model that predicts transcription factor–DNA affinities using structural and mechanical features of DNA, geometry, and flexibility, obtained from molecular dynamics simulations. Recently, leveraging the development of the ProNAB database, a deep learning method, DeePNAP,[61] has been developed using the sequence descriptors of proteins and nucleic acids.

Recent focus is shifting towards the prediction of protein–carbohydrate binding affinity. Initially, the prediction was made based on knowledge-based statistical potentials.[62] Later, the development of databases such as ProCaff and ProCarbDB accelerated the application of machine learning methods in this field. These methods mainly focus on the energetics of interactions[34] and interface contacts. Nguyen *et al.*[63] developed a prediction method based on the geometry and chemistry of interactions between the molecules using graph-based signatures. The performance of binding affinity predictions can be substantially improved through the integration of expanded high-quality experimental datasets and algorithms that better capture the structural and physicochemical determinants of molecular interactions.

4. Machine learning and AI-based methods for predicting the binding affinity change upon mutations

Mutations that alter the binding affinity of biomolecular complexes can profoundly impact cellular processes, disease phenotypes,

Table 2 Computational algorithms for predicting the binding affinity of protein complexes

Tools	Interaction type	Features	URL
PerSpect-EL[52]	Protein–protein	Persistent homology and physical properties	https://github.com/ExpectozJJ/PerSpect-Ensemble-Learning
PRODIGY[53]	Protein–protein	Network of inter-residue contacts and non-interacting surface	https://github.com/haddocking/prodigy
PPI-Affinity[54]	Protein–protein	Structure-based features (ProtDcal)	https://protdcal.zmb.uni-due.de/PPIAffinity
PPA-Pred[15]	Protein–protein	Sequence-based features	https://www.iitm.ac.in/bioinfo/PPA_Pred/
ISLAND[55]	Protein–protein	Protein sequence information using kernel representation	https://sites.google.com/view/wajidarshad/software
PIPR[16]	Protein–protein	Robust local features and contextualized information	https://github.com/muhaochen/seq_ppi
PPI-Graphomer[13]	Protein–protein	Sequence and structural features are extracted using ESM2 and ESM-IF1Graph transformer model	https://github.com/xiebaoshu058/PPI-Graphomer
DeepPPAPred[30]	Protein–protein	Features from the sequence information and predicted three-dimensional structures.	https://web.iitm.ac.in/bioinfo2/DeepPPAPred/index.html
ProAffinity-GNN[51]	Protein–protein	Protein language model and graph neural network (GNN) using structures	https://github.com/legendzzy/ProAffinity-GNN
ProBAN[18]	Protein–protein	Location of atoms and their abilities to participate in various types of interactions	https://github.com/EABogdanova/ProBAN
AREA-AFFINITY[49]	Protein–protein	Geometric characteristics such as area (both interface and surface areas)	https://affinity.cuhk.edu.cn/
MPA-Pred[31]	Protein–protein	Specific for membrane protein complexes, sequence and structure-based features	https://web.iitm.ac.in/bioinfo2/MPA-Pred/
DNAffinity[56]	Protein–DNA	Molecular dynamics simulation-based features	https://github.com/Jalbiti/DNAffinity
PreDBA[57]	Protein–DNA	An ensemble model using sequence and structural features of the protein and DNA	https://predba.denglab.org/
PDA-Pred[32]	Protein–DNA	Interaction features, volume and surface area of the interface, DNA base step parameters, and atom contacts	https://web.iitm.ac.in/bioinfo2/pdapred/
emPDBA[58]	Protein–DNA	Sequence, structure, and interface features of the complex and the individual partners	https://github.com/ChunhuaLiLab/emPDBA/
PredPRBA[59]	Protein–RNA	Interface hydrophobicity, hydration pattern, and change in the conformation due to binding	https://predprba.denglab.org/
PRA-Pred[33]	Protein–RNA	Contact-based features, interaction energies, RNA base step parameters, and hydrogen bonding	https://web.iitm.ac.in/bioinfo2/prapred/
PNAB[60]	Protein–nucleic acid	Physicochemical properties, protein and nucleic acid sequence-based features	https://pnab.denglab.org/
DeePNAP[61]	Protein–nucleic acid	Deep learning method utilizing sequence descriptor of proteins and nucleic acids	http://14.139.174.41:8080/
PCA-Pred[34]	Protein–carbohydrate	Structure-based features such as contact potentials, interaction energy, number of binding residues, and contacts	https://web.iitm.ac.in/bioinfo2/pcapred/
SPOT-Struc[62]	Protein–carbohydrate	Knowledge-based statistical potential	NA
CSM-carbohydrate[63]	Protein–carbohydrate	Information on both protein and carbohydrate complementarity, in terms of shape and chemistry, was captured using graph-based structural signatures	https://biosig.lab.uq.edu.au/csm_carbohydrate/

and therapeutic efficacy.[66] From disrupting protein–protein signaling cascades to modulating the specificity of protein–DNA or protein–carbohydrate interactions, even single amino acid substitutions at key interface residues can lead to substantial functional consequences. Accurate prediction of how mutations affect binding affinity ($\Delta\Delta G$) is therefore essential not only for understanding the molecular basis of diseases but also for guiding rational protein design and the development of precision therapeutics.

Before the advent of machine learning and AI-based approaches, predictions of mutational effects focused on physics-based and knowledge-based methods. Methods like FoldX[67] use empirical force fields to estimate free energy changes, while BeAtMuSiC,[68] BindProf,[69] and BindProfX[70] leverage structural, energetic, and evolutionary features, focusing on statistical potentials rather than purely data-driven learning approaches. These traditional approaches laid the groundwork for the development of AI-driven models. Table 3 summarizes key computational methods for predicting mutation-induced binding affinity changes, including both traditional and AI-based approaches.

The predictive accuracy of machine learning (ML) models for estimating binding affinity changes upon mutation depends on two main factors: the choice of features and the algorithmic architecture. Early efforts in this domain relied heavily on features derived from sequences or experimental structures. These features included physicochemical descriptors of mutated residues, changes in solvent-accessible surface area, hydrogen bonding patterns, and electrostatic potentials at the interface. Energetic terms, such as van der Waals contributions or binding free energy approximations obtained from empirical force fields (*e.g.*, FoldX,[67] Rosetta[84]), were often integrated to enhance biophysical interpretability. Sequence-based features, particularly those capturing evolutionary conservation

Table 3 Traditional and machine learning based methods for predicting binding affinity change upon mutations

Tools	Interaction type	Features	URL
BeAtMuSiC[68]	Protein–protein	Statistical potentials derived from structure	https://babylone.ulb.ac.be/beatmusic/index.php
BindProf[69]	Protein–protein	Interface structure profile, physics-based potentials, and sequence-based profile	https://zhanglab.ccmb.med.umich.edu/BindProf/
BindProfX[70]	Protein–protein	Interface profile and FoldX physics potential	https://zhanglab.ccmb.med.umich.edu/BindProfX/
MutaBind2[71]	Protein–protein	van der Waals energy, solvation energy, unfolding free energy, SASA, conservation score and interfacial contacts	https://lilab.jysw.suda.edu.cn/research/mutabind2/
iSEE[72]	Protein–protein	Interface structure profile, evolution and energy-based features	https://github.com/haddocking/iSee
mCSM-PPI2[73]	Protein–protein	Graph-based signature, evolutionary information, complex network metrics, and energetic terms	https://biosig.lab.uq.edu.au/mcsm_ppi2/
TopNetTree[74]	Protein–protein	Persistent homology-based topological descriptors and CNN-derived features	https://codeocean.com/capsule/2202829/tree/v1
SAAMBE-3D[75]	Protein–Protein	Knowledge-based features from the mutation site environment	https://compbio.clemson.edu/saambe_webserver/
GeoPPI[76]	Protein–protein	Geometric deep features from structure	https://github.com/Liuxg16/GeoPPI
DDMut-PPI[17]	Protein–protein	ProtT5 embeddings and interaction-type graph edges	https://biosig.lab.uq.edu.au/ddmut_ppi/
ProAffiMuSeq[7]	Protein–protein	Amino acid properties, PSSM, interface-specific indices and protein functional classes	https://web.iitm.ac.in/bioinfo2/proaffimuseq/
PANDA[77]	Protein–protein	Amino acid composition, conservation score, physicochemical properties	https://pandaaffinity.pythonanywhere.com/
SAAMBE-SEQ[78]	Protein–protein	Evolutionary, sequence, and physicochemical features of the mutation site	https://compbio.clemson.edu/saambe_webserver/indexSEQ.php
DeepPPAPredMut[20]	Protein–protein	Physicochemical, evolutionary, and graph-based features	https://web.iitm.ac.in/bioinfo2/DeepPPAPredMut/
MPA-MutPred[35]	Membrane protein–protein	Electrostatic interaction, SASA, conservation score and interfacial contacts	https://web.iitm.ac.in/bioinfo2/MPA-MutPred/
PremPRI[79]	Protein–RNA	Interface interactions and graph-based features	https://lilab.jysw.suda.edu.cn/research/PremPRI/.
PremPDI[80]	Protein–DNA	Molecular mechanics, statistical potentials and accessibility	https://lilab.jysw.suda.edu.cn/research/PremPDI/
SAMPDI-3Dv2[81]	Protein–DNA	Structural features and knowledge-based terms (protein and DNA)	https://compbio.clemson.edu/SAMPDI-3D/
mCSM–NA[82]	Protein–NA	Graph-based signatures utilizing the encoded amino acid residue	https://biosig.lab.uq.edu.au/mcsm_na/
PEMPNI[83]	Protein–NA	Energy-based and structural interface features, such as contacts and residue-nucleotide pairs	https://liulab.hzau.edu.cn/PEMPNI
PRA-Mut-Pred[36]	Protein–RNA	Structural, Energy-based and network-based features are utilized for the prediction using Support Vector Algorithm	https://web.iitm.ac.in/bioinfo2/pramutpred/
PCA-MutPred[37]	Protein–carbohydrate	Sequence and structure-based features using multiple linear regression techniques	https://web.iitm.ac.in/bioinfo2/pcamutpred

(*e.g.*, position-specific scoring matrices), provided an orthogonal source of information and proved especially useful in identifying mutation-sensitive hotspots at conserved interfaces.[7,71]

With the emergence of deep learning, the emphasis has shifted from manual feature engineering to data-driven representation learning. Structure-based models often employ convolutional neural networks (CNNs) to capture spatial and geometric patterns around mutation sites, incorporating both local and topological information.[74] More recently, graph neural networks (GNNs) have offered a more flexible framework for representing biomolecules as graphs, where nodes correspond to atoms or residues and edges capture interatomic interactions.[76] Transformer-based models, originally developed for natural language processing, have also been repurposed to learn contextual embeddings of protein and nucleic acid sequences. When pretrained on large-scale sequence databases, these models (*e.g.*, ProtBert, ProtT5,[85] ESM[86]) can infer structural and evolutionary constraints implicitly, enabling them to generalize to unseen mutations. Hybrid models that combine these sequence embeddings with structural features,[17] either explicitly or *via* attention mechanisms, have demonstrated improved generalization in cross-domain applications.

A key development in recent years has been the integration of structure prediction pipelines into affinity prediction workflows. Tools such as AlphaFold-Multimer[10] and RoseTTAFoldNA[11] are now commonly used to generate mutant complex models, which serve as inputs for downstream feature extraction. These approaches have enabled mutation scanning even in the absence of high-resolution structural data, broadening the applicability of ML methods to underrepresented systems such as membrane protein–protein or protein–glycan complexes.

Despite substantial progress, challenges remain. Available datasets are limited in size and diversity, with an overrepresentation of mutations that cause minimal changes in binding affinity. Structural complexity, such as nucleic acid flexibility and glycan dynamics, is typically underrepresented in models, which often assume static interfaces. These limitations hinder generalization, especially across diverse mutation types and

binding mechanisms, underscoring the need for broader datasets and more adaptable modeling strategies.

5. Applications to design protein complexes

The computational design of protein complexes with tailored binding affinities is a rapidly growing field, powered by advances in structure prediction, scoring functions, and AI-guided mutational scanning.[87] The computational design of protein–protein complexes aims to engineer or optimize binding interfaces to achieve desired affinities, specificities, or functional properties. Recent advancements in structural prediction tools combined with affinity prediction algorithms have enabled more reliable modeling and redesign of PPIs. Key strategies include interface residue scanning, backbone flexibility modeling, and machine learning-based affinity ranking.[88] Binding affinity plays a decisive role in evaluating design success, as complexes with suboptimal affinity may fail to form under physiological conditions. Several studies have demonstrated that tuning interfacial hydrophobicity, electrostatic complementarity, and hydrogen bonding patterns can significantly enhance binding.[89] In addition, approaches that integrate molecular dynamics simulations and free energy perturbation analyses have elucidated how subtle changes in interface polarity or residue packing can modulate association energetics. Recent developments in deep learning-based energy models further enable residue-level optimization, capturing long-range interactions that traditional scoring functions may overlook.[90] A recent study demonstrates that interface design guided by computational binding affinity predictions can yield novel, hyperstable binders with nanomolar to picomolar affinities, as confirmed by structural and biophysical validation.[91] Applications span therapeutic protein development, synthetic biology, and enzyme engineering, where controlling PPI strength is often critical to function.

Protein–DNA/RNA complexes are pivotal for gene regulation, epigenetics, and cellular signaling, and accurate three-dimensional structure prediction of these complexes enhances our understanding of their atomic-level recognition, molecular functions, and binding affinities. Further, this understanding holds significant potential in computer-aided and structure-based drug design. Protein–nucleic acid complexes are mainly generated using template-based docking. In addition, *ab initio* and machine learning approaches are also used for modeling their three-dimensional structures.[92] Unlike protein–protein interactions, rational design of protein–DNA/RNA complexes remains largely unachieved due to nucleic acid flexibility and conformational heterogeneity (see Section 6 for further discussion).

Protein–carbohydrate interactions play a major role in inflammation, cell proliferation, differentiation, aggregation, signal transduction, host–pathogen recognition, and protein structure stabilization. Computational methods enable the study of diverse carbohydrate systems, providing insights into their structures, dynamics, and interactions.[27] However, modeling protein–carbohydrate complexes remains challenging due to low affinity, multivalency, and structural heterogeneity, as many carbohydrate-binding proteins, including lectins and adhesins, achieve specificity by binding multiple identical glycoside units arranged in distinct patterns.

In essence, computational design of protein–protein complexes has achieved notable success. In contrast, designing protein–DNA/RNA and protein–carbohydrate complexes remains highly challenging due to conformational flexibility, multivalency, and structural heterogeneity. The underlying difficulties and potential strategies to overcome these challenges are discussed in detail in the next section, highlighting directions for future research in the field.

6. Future perspectives

The design of biomolecular complexes guided by binding affinity remains a key challenge in structural biology and therapeutic development. Recent advances in structure prediction have transformed our ability to model large macromolecular assemblies across diverse interaction types, including protein–protein, protein–DNA/RNA, and protein–carbohydrate complexes. This opens up new opportunities to explore binding affinity landscapes across sequence and structure space. However, leveraging these structural models for quantitative affinity prediction and rational design demands further progress on multiple fronts.

A major limitation in current approaches is the quality and diversity of available training data. For protein–protein systems, while curated databases such as SKEMPI and PROXiMATE have enabled affinity prediction and mutation effect estimation, biases in sequence diversity and mutation type, especially over-representation of alanine scanning, remain a concern. To advance predictive modeling, there is a growing need for machine-learning-grade datasets[93] that are large in size, diverse, well-annotated, standardized, and curated to capture the biochemical and structural complexity necessary for training robust and generalizable models. Such a volume of data will become increasingly feasible with advances in high-throughput experimental methods, such as deep mutational scanning and multiplexed binding assays, which enable systematic and scalable measurement of binding affinities. Heyne *et al.*[94] developed a novel high-throughput approach for obtaining changes in binding free energy data for thousands of protein mutants in a single experiment and opening a new way for studies of mutation effects in PPIs. The approach combines yeast surface display, deep sequencing, and data normalization, producing affinity measurements comparable to traditional low-throughput methods.

To enable robust predictive and design-driven applications, future computational tools must go beyond static sequence and structure, incorporating key determinants such as post-translational modifications, intrinsic disorder, allosteric regulation, and dynamic conformational states. Bridging these aspects will be critical for advancing from affinity prediction

to rational design of synthetic interfaces, therapeutic antibodies, and other engineered biomolecular assemblies. The design of multi-specific binders, interface stabilization, and re-engineering of host–pathogen interactions (*e.g.*, SARS-CoV-2 spike–ACE2, or broadly neutralizing antibodies) are promising areas where predictive modeling could yield tangible impact.

For protein–nucleic acid complexes, challenges are more acute. Homology/template-based docking methods are limited by available structural templates, *ab initio* approaches are computationally intensive, and deep learning-based predictions have yet to achieve robust performance in this space. Significant advances have been made in enhancing the prediction quality through the selection of optimum parameters and large datasets.[92,95] However, performance remains limited, and further developments are needed to accurately predict protein–nucleic acid complex structures, as evidenced from the recent CASP experiments.[96]

Protein–nucleic acid interactions have been characterized by assessing the relationship between three-dimensional structural features and binding affinity. Despite the availability of a few methods for predicting their binding affinity, there exists a substantial demand for improving their performance uniformly across different types of complexes, considering variations in structure and function. Further, there is an opportunity to develop methods that can predict the binding affinity from the protein and nucleic acid sequences directly. Improved understanding of these interactions could enable the design of high-affinity aptamers to selectively disrupt the complex formation, offering potential therapeutic opportunities, for example, by targeting long non-coding RNAs (lncRNAs) and large intergenic non-coding RNAs (lincRNAs) implicated in the epigenetic regulation of cancer-related gene expression.[97,98] Moreover, many protein–nucleic acid complexes involve multiple protein subunits (*e.g.*, dimers or trimers) binding to nucleic acids, which must be appropriately accounted for in prediction models. Existing methods predominantly focus on single amino acid mutations, often overlooking the impact of mutations on nucleic acids. Accurately predicting the impact of mutations, including disease-associated variants at protein–nucleic acid interfaces, on binding affinity is critical for understanding molecular mechanisms and guiding drug discovery.

Protein–carbohydrate interactions present a distinct set of challenges compared to protein–protein or protein–nucleic acid complexes. While tools like AlphaFold3 now enable structural modeling of protein–carbohydrate assemblies, the intrinsic flexibility, multivalency, and structural heterogeneity of glycans complicate both binding prediction and rational design. On biological surfaces, glycans are rarely isolated; instead, they are organized into glycolipid- and glycoprotein-derived clusters that form distinct glyco-surface patches. Together, these patches create a complex 3D glyco landscape, in which specific recognition motifs—termed glycotopes—govern selective binding. Accurately modeling these features, including the multivalent and water-mediated nature of glycan interactions,[99] is critical for the successful design of protein–carbohydrate complexes.

Future directions in computational protein–carbohydrate interaction design are increasingly being shaped by AI-driven tools. Deep learning models like GlyNet[100] and GlyBERT,[101] combined with AlphaFold-like structure prediction methods, are improving the modeling of flexible glycan-binding interfaces. Expanding structural databases with glycan-bound complexes through experiments or AI predictions, integrating physics-based and graph neural network approaches,[102,103] and employing deep generative models could accelerate the design of synthetic lectins, glycan sensors, and glycan-targeting proteins with tailored affinity and specificity. Iterative workflows that combine computational modeling with experimental feedback (*e.g.*, glycan arrays or SPR) will further refine predictive accuracy. These strategies are essential for therapeutic applications, enabling the rational engineering of glycan-binding proteins that support the development of antibodies, vaccines, and inhibitors targeting lectins, carbohydrate-active enzymes, glycosaminoglycans, and other glycan-modified biomolecules. Computational workflows that account for the organization of glycans on cell surfaces, their interactions within the extended carbohydrate layer (glyco-canopy), and their multivalent binding will be key to enabling predictive and design-oriented strategies in glyco-engineering.

Looking ahead, the design of biomolecular complexes with tailored binding affinity will increasingly rely on the integration of structure prediction, mechanistic modeling, and data-driven inference. Advances in deep learning models have enabled high-resolution structural modeling across diverse interaction types, yet leveraging these structures for accurate affinity prediction and rational design remains challenging. Progress will depend on the development of standardized, diverse, and well-annotated datasets, particularly for underrepresented systems such as membrane protein–protein and protein–carbohydrate complexes. Incorporating biologically relevant features, such as conformational flexibility, post-translational modifications, and cellular context, will improve model robustness and applicability. Coupling predictive tools with high-throughput experimental platforms will enable iterative, feedback-driven design workflows. Also, moving beyond static binary interactions toward more complex assemblies will be critical for advancing our ability to engineer functional biomolecular systems.

Ultimately, future progress will depend on unifying structural, thermodynamic, and mutational data into interoperable, diverse, and openly accessible platforms. By addressing current methodological and data limitations, the community can develop more robust models of affinity and expand the reach of computational design to novel molecular functions and therapeutic strategies.

7. Conclusions

Binding affinity is a central determinant of biomolecular recognition and function, and understanding its molecular basis is crucial for the rational design of protein complexes. In this review, we examined the importance of binding affinity in the

computational design landscape, from early data-driven modeling efforts to the integration of cutting-edge deep learning approaches. A key focus has been the role of high-quality experimental data, cataloged in curated databases spanning protein–protein, protein–DNA/RNA, and protein–carbohydrate interactions, in enabling both predictive modeling and design applications.

We discussed the progression of affinity prediction methods, highlighting both classical physics-based approaches and the growing impact of machine learning and deep learning models. These tools have been instrumental in predicting the effects of mutations, guiding interface redesign, and informing therapeutic engineering, particularly in antibody–antigen and host–pathogen systems such as SARS-CoV-2. To support these efforts, databases, including those developed by our group, broaden the coverage of affinity data and contribute to a more inclusive and diverse modeling foundation. Looking ahead, the integration of high-resolution structure prediction tools, high-throughput mutational scanning, and context-aware affinity models will enable more realistic and functionally relevant designs. Unified platforms that combine thermodynamic, structural, and functional annotations will be essential to advance modeling capabilities across biomolecular interaction types.

In summary, advancing the computational design of protein complexes will depend not only on algorithmic innovation but also on the quality and diversity of underlying data. By prioritizing integrative, high-fidelity datasets and refining model evaluation strategies, the field is poised to translate affinity prediction into more reliable and application-driven biomolecular design.

Author contributions

MMG conceived and supervised the review. FR, KH, NRS, and RN conducted the literature survey and drafted the manuscript. FR, KH, and NRS contributed to the preparation and editing of the graphical abstract. All authors discussed and revised the manuscript and approved the final version.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this review.

Acknowledgements

We are thankful to IIT Madras for offering infrastructural facilities. The work is partially supported by the Department of Biotechnology, Government of India, to MMG (BT/PR39164/BID/7/965/2020).

Notes and references

1 T. Pawson and P. Nash, *Science*, 2003, **300**, 445–452.
2 B. Alberts, *Mol. Biol. Cell*, 2010, **21**, 3785.
3 Y. I. Network, *Science*, 2008, **1158684**, 322.
4 G. Schreiber and A. E. Keating, *Curr. Opin. Struct. Biol.*, 2011, **21**, 50–61.
5 I. Ezkurdia, L. Bartoli, P. Fariselli, R. Casadio, A. Valencia and M. L. Tress, *Briefings Bioinf.*, 2009, **10**, 233–246.
6 L. Maveyraud and L. Mourey, *Molecules*, 2020, **25**, 1030.
7 S. Jemimah, M. Sekijima and M. M. Gromiha, *Bioinformatics*, 2020, **36**, 1725–1730.
8 B. A. Shoemaker and A. R. Panchenko, *PLoS Comput. Biol.*, 2007, **3**, e42.
9 J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli and D. Hassabis, *Nature*, 2021, **596**, 583–589.
10 R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper and D. Hassabis, *bioRxiv*, 2021, preprint, DOI: 10.1101/2021.10.04.463034.
11 M. Baek, R. McHugh, I. Anishchenko, H. Jiang, D. Baker and F. DiMaio, *Nat. Methods*, 2024, **21**, 117–121.
12 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard and J. Bambrick, *Nature*, 2024, **630**, 493–500.
13 J. Xie, Y. Zhang, Z. Wang, X. Jin, X. Lu, S. Ge and X. Min, *BMC Bioinf.*, 2025, **26**, 116.
14 N. Homeyer and H. Gohlke, *Mol. Inf.*, 2012, **31**, 114–122.
15 K. Yugandhar and M. M. Gromiha, *Bioinformatics*, 2014, **30**, 3583–3589.
16 M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo and W. Wang, *Bioinformatics*, 2019, **35**, i305–i314.
17 Y. Zhou, Y. Myung, C. H. Rodrigues and D. B. Ascher, *Nucleic Acids Res.*, 2024, **52**, W207–W214.
18 E. A. Bogdanova and V. N. Novoseletsky, *Proteins: Struct., Funct., Bioinf.*, 2024, **92**, 1127–1136.
19 M. Réau, N. Renaud, L. C. Xue and A. M. Bonvin, *Bioinformatics*, 2023, **39**, btac759.
20 R. Nikam, S. Jemimah and M. M. Gromiha, *Bioinformatics*, 2024, **40**, btae309.
21 J. Jankauskaite, B. Jiménez-García, J. Dapkunas, J. Fernández-Recio and I. H. Moal, *Bioinformatics*, 2019, **35**, 462–469.
22 Z. Liu, Y. Li, L. Han, J. Li, J. Liu, Z. Zhao, W. Nie, Y. Liu and R. Wang, *Bioinformatics*, 2015, **31**, 405–412.
23 S. Jemimah, K. Yugandhar and M. Michael Gromiha, *Bioinformatics*, 2017, **33**, 2787–2788.
24 F. Ridha, A. Kulandaisamy and M. M. Gromiha, *J. Mol. Biol.*, 2023, **435**, 167870.
25 K. Harini, A. Srivastava, A. Kulandaisamy and M. M. Gromiha, *Nucleic Acids Res.*, 2022, **50**, D1528–D1534.
26 N. R. Siva Shanmugam, J. Jino Blessy, K. Veluraja and M. Michael Gromiha, *Bioinformatics*, 2020, **36**, 3615–3617.
27 S. Pérez, *Advances in Carbohydrate Chemistry and Biochemistry*, Elsevier, 2023, vol. 83, pp. 133–149.
28 N. R. Siva Shanmugam, J. Jino Blessy, K. Veluraja and M. M. Gromiha, *Briefings Bioinf.*, 2021, **22**, bbaa319.
29 P. L. Kastritis and A. M. Bonvin, *J. R. Soc., Interface*, 2012, **10**, 20120835.
30 R. Nikam, K. Yugandhar and M. M. Gromiha, *Biochim. Biophys. Acta, Proteins Proteomics*, 2023, **1871**, 140948.
31 F. Ridha and M. M. Gromiha, *Proteins: Struct., Funct., Bioinf.*, 2024, **92**, 499–508.
32 K. Harini, D. Kihara and M. M. Gromiha, *Methods*, 2023, **213**, 10–17.
33 K. Harini, M. Sekijima and M. M. Gromiha, *Int. J. Biol. Macromol.*, 2024, **259**, 129490.
34 N. R. Siva Shanmugam, J. Jino Blessy, K. Veluraja and M. M. Gromiha, *Briefings Bioinf.*, 2021, **22**, bbaa319.

35 F. Ridha and M. M. Gromiha, *Briefings Bioinf.*, 2024, **25**, bbae598.
36 K. Harini, M. Sekijima and M. M. Gromiha, *J. Chem. Inf. Model.*, 2025, **65**, 1605–1614.
37 N. R. Siva Shanmugam, K. Veluraja and M. M. Gromiha, *J. Mol. Biol.*, 2022, **434**, 167526.
38 T. Vreven, I. H. Moal, A. Vangone, B. G. Pierce, P. L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P. A. Bates and J. Fernandez-Recio, *J. Mol. Biol.*, 2015, **427**, 3031–3041.
39 S. Sirin, J. R. Apgar, E. M. Bennett and A. E. Keating, *Protein Sci.*, 2016, **25**, 393–409.
40 J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi and C. M. Deane, *Nucleic Acids Res.*, 2014, **42**, D1140–D1146.
41 T. Borrman, J. Cimons, M. Cosiano, M. Purcaro, B. G. Pierce, B. M. Baker and Z. Weng, *Proteins: Struct., Funct., Bioinf.*, 2017, **85**, 908–916.
42 H. Liu, P. Chen, X. Zhai, K.-G. Huo, S. Zhou, L. Han and G. Fan, *Sci. Data*, 2024, **11**, 1316.
43 P. Prabakaran, J. An, M. M. Gromiha, S. Selvaraj, H. Uedaira, H. Kono and A. Sarai, *Bioinformatics*, 2001, **17**, 1027–1034.
44 L. Liu, Y. Xiong, H. Gao, D.-Q. Wei, J. C. Mitchell and X. Zhu, *Database*, 2018, **2018**, bay034.
45 L.-C. Mei, G.-F. Hao and G.-F. Yang, *Trends Biotechnol.*, 2023, **41**, 140–143.
46 N. R. Siva Shanmugam, A. Kulandaisamy, K. Veluraja and M. M. Gromiha, *Glycobiology*, 2024, **34**, cwae011.
47 L. Copoiu, P. H. Torres, D. B. Ascher, T. L. Blundell and S. Malhotra, *Nucleic Acids Res.*, 2020, **48**, D368–D375.
48 M. M. Gromiha and K. Harini, *Trends Biotechnol.*, 2023, **41**, 988–989.
49 Y. X. Yang, J. Y. Huang, P. Wang and B. T. Zhu, *J. Chem. Inf. Model.*, 2023, **63**, 3230–3237.
50 A. Vangone and A. M. Bonvin, *eLife*, 2015, **4**, e07454.
51 Z. Zhou, Y. Yin, H. Han, Y. Jia, J. H. Koh, A. W.-K. Kong and Y. Mu, *J. Chem. Inf. Model.*, 2024, **64**, 8796–8808.
52 J. Wee and K. Xia, *Briefings Bioinf.*, 2022, **23**, bbac024.
53 L. C. Xue, J. o P. Rodrigues, P. L. Kastritis, A. M. Bonvin and A. Vangone, *Bioinformatics*, 2016, **32**, 3676–3678.
54 S. Romero-Molina, Y. B. Ruiz-Blanco, J. Mieres-Perez, M. Harms, J. Munch, M. Ehrmann and E. Sanchez-Garcia, *J. Proteome Res.*, 2022, **21**, 1829–1841.
55 W. A. Abbasi, A. Yaseen, F. U. Hassan, S. Andleeb and F. U. A. A. Minhas, *BioData Min.*, 2020, **13**, 20.
56 S. Barissi, A. Sala, M. Wieczór, F. Battistini and M. Orozco, *Nucleic Acids Res.*, 2022, **50**, 9105–9114.
57 W. Yang and L. Deng, *Sci. Rep.*, 2020, **10**, 1278.
58 S. Yang, W. Gong, T. Zhou, X. Sun, L. Chen, W. Zhou and C. Li, *Briefings Bioinf.*, 2023, **24**, bbad192.
59 L. Deng, W. Yang and H. Liu, *Front. genet*, 2019, **10**, 637.
60 W. Yang and L. Deng, 2019.
61 U. Pandey, S. M. Behara, S. Sharma, R. S. Patil, S. Nambiar, D. Koner and H. Bhukya, *J. Chem. Inf. Model.*, 2024, **64**, 1806–1815.
62 T. B. Nguyen, D. E. Pires and D. B. Ascher, *Briefings Bioinf.*, 2022, **23**, bbab512.
63 H. Zhao, Y. Yang, M. von Itzstein and Y. Zhou, *J. Comput. Chem.*, 2014, **35**, 2177–2183.
64 P. Pant, A. Pathak and B. Jayaram, *J. Biomol. Struct. Dyn.*, 2023, **41**, 4040–4047.
65 D. Vuzman and Y. Levy, *Mol. BioSyst.*, 2012, **8**, 47–57.
66 M. M. Gromiha, M. Pandey, A. Kulandaisamy, D. Sharma and F. Ridha, *Comput. Biol. Med.*, 2025, **185**, 109510.
67 J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau and L. Serrano, *Nucleic Acids Res.*, 2005, **33**, W382–W388.
68 Y. Dehouck, J. M. Kwasigroch, M. Rooman and D. Gilis, *Nucleic Acids Res.*, 2013, **41**, W333–W339.
69 J. R. Brender and Y. Zhang, *PLoS Comput. Biol.*, 2015, **11**, e1004494.
70 P. Xiong, C. Zhang, W. Zheng and Y. Zhang, *J. Mol. Biol.*, 2017, **429**, 426–434.
71 N. Zhang, Y. Chen, H. Lu, F. Zhao, R. V. Alvarez, A. Goncearenco, A. R. Panchenko and M. Li, *iScience*, 2020, **23**, 100939.
72 C. Geng, A. Vangone, G. E. Folkers, L. C. Xue and A. M. Bonvin, *Proteins: Struct., Funct., Bioinf.*, 2019, **87**, 110–119.
73 C. H. Rodrigues, Y. Myung, D. E. Pires and D. B. Ascher, *Nucleic Acids Res.*, 2019, **47**, W338–W344.
74 M. Wang, Z. Cang and G.-W. Wei, *Nat. Mach. Intell.*, 2020, **2**, 116–123.
75 S. Pahari, G. Li, A. K. Murthy, S. Liang, R. Fragoza, H. Yu and E. Alexov, *Int. J. Mol. Sci.*, 2020, **21**, 2563.
76 X. Liu, Y. Luo, P. Li, S. Song and J. Peng, *PLoS Comput. Biol.*, 2021, **17**, e1009284.
77 W. A. Abbasi, S. A. Abbas and S. Andleeb, *J. Bioinf. Comput. Biol.*, 2021, **19**, 2150015.
78 G. Li, S. Pahari, A. K. Murthy, S. Liang, R. Fragoza, H. Yu and E. Alexov, *Bioinformatics*, 2021, **37**, 992–999.
79 N. Zhang, H. Lu, Y. Chen, Z. Zhu, Q. Yang, S. Wang and M. Li, *Int. J. Mol. Sci.*, 2020, **21**, 5560.
80 N. Zhang, Y. Chen, F. Zhao, Q. Yang, F. L. Simonetti and M. Li, *PLoS Comput. Biol.*, 2018, **14**, e1006615.
81 P. Rimal, S. K. Paul, S. K. Panday and E. Alexov, *Genes*, 2025, **16**, 101.
82 D. E. Pires and D. B. Ascher, *Nucleic Acids Res.*, 2017, **45**, W241–W246.
83 Y. Jiang, H.-F. Liu and R. Liu, *PLoS Comput. Biol.*, 2021, **17**, e1008951.
84 R. F. Alford, A. Leaver-Fay, J. R. Jeliazkov, M. J. O'Meara, F. P. DiMaio, H. Park, M. V. Shapovalov, P. D. Renfrew, V. K. Mulligan, K. Kappel, J. W. Labonte, M. S. Pacella, R. Bonneau, P. Bradley, R. L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme and J. J. Gray, *J. Chem. Theory Comput.*, 2017, **13**, 3031–3048.
85 A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik and B. Rost, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022, **44**, 7112–7127.
86 A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick and J. Ma, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2016239118.
87 P.-S. Huang, S. E. Boyken and D. Baker, *Nature*, 2016, **537**, 320–327.
88 E.-M. Strauch, S. J. Fleishman and D. Baker, *Proc. Natl. Acad. Sci. U. S. A.*, 2014, **111**, 675–680.
89 S. J. Fleishman, T. A. Whitehead, D. C. Ekiert, C. Dreyfus, J. E. Corn, E.-M. Strauch, I. A. Wilson and D. Baker, *Science*, 2011, **332**, 816–821.
90 J. Cui, S. Yang, L. Yi, Q. Xi, D. Yang and Y. Zuo, *BioData Min.*, 2025, **18**, 43.
91 L. Cao, B. Coventry, I. Goreshnik, B. Huang, W. Sheffler, J. S. Park, K. M. Jude, I. Marković, R. U. Kadam and K. H. Verschueren, *Nature*, 2022, **605**, 551–560.
92 M. M. Gromiha and K. Harini, *Curr. Opin. Struct. Biol.*, 2025, **90**, 102955.
93 A. M. Hummer, C. Schneider, L. Chinery and C. M. Deane, *Nat. Comput. Sci.*, 2025, 1–13.
94 M. Heyne, N. Papo and J. M. Shifman, *Nat. Commun.*, 2020, **11**, 297.
95 C. Zeng, Y. Jian, C. Zhuo, A. Li, C. Zeng and Y. Zhao, *Phys. Chem. Chem. Phys.*, 2024, **26**, 130–143.
96 W. Wang, Y. Luo, Z. Peng and J. Yang, *Proteins:Struct., Funct., Bioinf.*, 2025.
97 J. S. Mattick, P. P. Amaral, P. Carninci, S. Carpenter, H. Y. Chang, L.-L. Chen, R. Chen, C. Dean, M. E. Dinger and K. A. Fitzgerald, *Nat. Rev. Mol. Cell Biol.*, 2023, **24**, 430–447.
98 J. Cao, *Biol. Proced. Online*, 2014, **16**, 42.
99 S. Shin, M. L. Mugnai and D. Thirumalai, *J. Am. Chem. Soc.*, 2025, **147**, 17448–17458.
100 E. J. Carpenter, S. Seth, N. Yue, R. Greiner and R. Derda, *Chem. Sci.*, 2022, **13**, 6669–6686.
101 D. E. Mattox and C. Bailey-Kellogg, *PLoS Comput. Biol.*, 2021, **17**, e1009470.
102 R. Burkholz, J. Quackenbush and D. Bojar, *Cell Rep.*, 2021, **35**, 109251.
103 Y. Luo and F. Parmeggiani, *bioRxiv*, 2025, preprint, DOI: **10.1101/2025.02.27.640667**.

