Index

Note: Page numbers in *italics* and **bold** refers to figures and tables respectively.

a

ab initio modeling 185
ab initio protein design 165, 167
ab initio protein structure prediction method 16, 29, 85
accelerated molecular dynamics (aMD) 24, 26
acetyl (ACE) 122
affinity purification coupled with mass spectrometry (AP-MS) 129, 208
allosteric changes 16, 30
AlphaFold 3, 6, 8, 20, 21, 134–135, 139, 165, 176, 180, 187, 209, 211–215, 217
  architecture of 44–47, *45*
  labelled and unlabelled data 48
  network interpretation 48
AlphaFold2 53–54, 113, 150, 153, 157, 168, 176
AlphaFold3 176
AlphaFold-Multimer 131, 208
AlphaFold Protein Structure Database 178
Amber ff14SB 115
AMBER force field 23, 87–88, 135
amino acid
  mutations 57, 60–62
  sequence 1, 4, 15, *17, 86*
Anfinsen's dogma 168
angle-bending interactions 22
AnnoPRO 162
antimicrobial peptide (AMP) 194
artificial intelligence (AI) 113, 131, 150, 165, 198, 210
  AI-augmented experimental techniques 217–220
  AlphaFold from Google 43–48
  based methods 160–161
  ESM and ESMFold from Meta 48–54
  in protein prediction 3
  quantum mechanics calculations 7
  self-supervision 49
  transformative role 210–213
Asilomar Conference on Recombinant DNA 191
atomic interactions 6, 95
ATP binding 153
attention mechanisms 169–170, 172

b

backbone structure model 47, *47*
back propagation, error minimization 99
Basic Local Alignment Search Tool (BLAST) 150, 153, 155–156
Bayesian methods 140
Bayesian networks 136
Bayesian Ridge (BR) 66
benchmark datasets 38
Berendsen thermostat 88

Machine Learning in Protein Science: Efficient Prediction of Protein Structures and Properties, First Edition. Jinjin Li and Yanqiang Han.
© 2026 Wiley-VCH GmbH. Published 2026 by Wiley-VCH GmbH.

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 224 — #2

Bidirectional Encoder Representations from Transformers (BERT) 48, 138
binding affinity prediction, ML-based 7
BioBERT 138
biochemical assays 155
BioCreative challenges 138
BioGRID database 130, 137, 154
bioinformatics 179
biological datasets, for protein modeling 7
biological processes 149, 152
biophysical assays 219
biotechnology, advances through protein design 11
"black-box" models 36, 38, 182, 191
Blast2GO 158
B3LYP exchange-correlation 87
bonded interactions 22
bond stretching 22–23

c

calorimetry 58, 207
CAMEO structure 52–54, *52*
cellular components 149, 152
cellular metabolism 1
CHARMM force field 23–24, 27
chemical denaturants 58
chromatin immunoprecipitation (ChIP) 219
circular dichroism (CD) 58
Class, Architecture, Topology, Homology (CATH) 150, 153, 157
cloud-based platforms 182
ClusPro 133
clustered tree regression (CTR) 61–62, *62, 64,* 212
  data distribution and cleaning 62–63
  data preparation 68
  experimental setup 69
  feature engineering 63–65
  feature extraction 68–69
  model training 65
  performance of 65–68
co-evolutionary analysis 132
co-immunoprecipitation (Co-IP) 129, 212
command line interface (CLI) 62
community detection methods 136
computational approaches, in protein research 8
computational biology 179
computational chemistry 111
computational docking simulations 141
computational protein biology, in machine learning 3
computational protein design, ESM in 197
computational protein science 15
conjugate caps 88, 100, 103
contrastive learning 190
convolutional neural networks (CNNs), 6, 34, 131–132, 140, 189. *see also* graph neural networks (GNNs)
Coulomb's Law 23
coupled-cluster theory 81, 99
CRISPR-based screening 162
Critical Assessment of protein Function Annotation (CAFA) 152
Critical Assessment of Protein Structure Prediction (CASP) 43, 187
  CASP14 43, 53–54, 168
cryo-electron microscopy (cryo-EM) 1, 6, 7–8, 11, 150, 156, 170, 185, 207, 211, 217, 219

d

data augmentation methods 48, 122, 169
data integration, in ML-based protein models 8
dataset bias 98
DeepGO 150, 160, 177, 209, 215
deep learning
  assisted quantum mechanical calculations 11
  full-system quantum modeling 77
  models 131, 140, 150, 174
  in protein prediction 20
  in protein structure 3
DeepMind 165, 168, 176, 180, 187
deep neural networks (DNNs) 113
  application of 6
  DNN–QM hybrid models 80
  GFCC method 84–95
  molecular calculations 81–83, *82*
  in protein prediction 3
DeepPotential 165, *166*

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 225 — #3

DeepPPI 131, 208
denaturant titration, protein unfolding studies 58
density functional theory (DFT) 20, 29, 78, 81, 113
differential scanning calorimetry (DSC) 58
direct coupling analysis (DCA) 132
directed evolution 197
disease mechanisms 17
  protein mutations in 4
distance matrices 38
DNA binding 153
docking algorithms 188
domain-motif recognition techniques 141
double-hybrid DFT calculations 114, *116*
dropout techniques 99, 122
drug discovery
  deep learning–based virtual screening 79
  ML applications in 7, 34–35
drug–protein interaction *3*
drug–receptor interaction 30
drug resistance, protein structure role 57
dynamic graph neural networks (DGNNs) 172

e

early stopping techniques 99
EASE-MM method 60
EF-hand domain 132
electron correlation 77
electronic structure of proteins 33
electrostatic calculations 27
electrostatic interactions 23, 58
energy barriers 30
energy function of neural networks for protein design 173–175
energy minimization techniques 165
energy reconstruction 98–101
engineered proteins 216
enhanced sampling techniques 59
environmental applications of proteins 198
enzymatic assays 218
enzyme catalysis 19, 23, 27, 30, 31, 33
ESMFold model 53–54, *53*
Evoformer block architecture 45–47, *46*
evolutionary scale modeling (ESM) 48–51, 187, 189, 209, 213
  decoding protein interactions using 195–201
  ESM-1b model, development of 52
  ESM-2 model *52,* 53, 196
  ESM3 196, 197
  future directions 198–201
  protein design, applications in 196–198
  technical foundations 195–196
experimental approaches 155, 219
  integration with 198
experimental biophysics 179
explainable AI (XAI) 38, 93, 105, 182
explicit solvent models 33

f

fast Fourier transform (FFT) algorithm 133
feature attribution methods 170
fine-tuning 114
FireProt dataset 62, *63*
FireProtDB 68
folding free energy 57–59
folding stability 174
FoldX method 59
force fields (FFs) 112
  force field-based calculation, on protein systems 21–28
fragmental algorithm 115, 121, *123*
fragment-based double-hybrid DFT 115
free energy calculations 113
free energy change (ΔΔG), predictions protein folding 60–66, *66*
full quantum mechanics (FQM) calculations 11, 111, 112, 120–121, 124
  classical methods to TL in protein 112–113
  GFCC method, DNN-based 84–95
  molecular calculations, DNNs for 81–83
  two-body molecular fractionation with conjugate caps 95–106
Fundamental AI Research (FAIR) Protein 196

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 226 — #4

g

Gaussian 16 software package 87
GDF8 protein sequences 161, *161*
GDF11 protein sequences 161, *161*
gene knockout experiments 4
Gene Ontology (GO) model 138, 149–150, 159, 166–167, 209
Generalized Born (GB) model 25
generalized molecular fractionation with conjugate caps (GFCC) method 84–95
Generative Adversarial Networks (GANs) 168, 174, 178, 209
generative models 168–169, 209
  for novel proteins 197
generative NNs 174
gene regulatory networks 165
genetic mutations, impact on protein function 10
genetic studies 155
genetic variants, ML-based prediction methods 59
genome-wide association studies (GWAS) 215
genomic data 149, 154–155
genomics 199
Gibbs free energy 57
Global Distance Test (GDT) score 48, 176
glycan masking 176
GNAT 137
Gradient Boosting (GB) 66, 140
Graph Attention Networks (GATs) 136
graph-based GANs 171
Graph Convolutional Networks (GCNs) 136
graph explainability methods 172
graphical user interface (GUI) 62
graphics processing units (GPUs) 83, 189
graph neural networks (GNNs) 7, 20, 35, 38, 131, 139, 171, 179–182, 208, 212
  energy function of neural networks for protein design 173–175
  extracting features, process of *173*
  for flexible protein design 171–182
  ML and AI-based methods 177–179
  ML models and technologies 176
Green Fluorescent Protein (GFP) mutants 169, *170*
GROMACS software 24, 88, 89, 135
"guilt by association" principle 166

h

HADDOCK 133
Hartree–Fock (HF) theory 29
Hidden Markov Model for sequence alignment (HMMER) 150, 153, 155–156
hidden Markov models (HMMs) 132, 186
high-dimensional data, protein modeling 77
higher-order interactions 97–98, 103
high-throughput experimental techniques 161–162, 182
high-throughput screening (HTS) techniques 6, 11, 21, 59, 79, 89, 90, 96, 105, 219
homology-based annotation 150, 153, 156, 209
homology-based prediction 132
homology modeling 16, 134–135, 185
Hooke's Law 22
hub proteins 135
hybrid approaches 5, 112, 139–140, *140,* 162–163
hybrid quantum-classical methods 23

i

I-Mutant2.0 method 59
inductive transfer learning (ITL) 114
inductive transfer learning force field (ITLFF) 112–113, *115,* 118–120
  implications 116–118
  methods 114–116
  on protein 5 k26 *118*
industrial applications of proteins 197
Inferred from Direct Assay (IDA) 150
Inferred from Electronic Annotation (IEA) 150
IntAct database 130, 137, 154

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 227 — #5

InterPro protein 17, 132
Invariant Point Attention (IPA) 47

k

Kernel Ridge (KR) 66
K-means clustering algorithm 61, 64
knowledge-based methods 59
knowledge distillation 191

l

labelled data, supervised learning on 48
language model 52
large language models (LLMs) 186, 209, 213
  applications of PLMs 192–194
  decoding protein interactions using ESM 195–201
  inputting protein sequences *187*
  protein language models, methodology and training of 189–191
  in protein systems 186–188
  visualization of LLMs tools *214*
lattice models 79
LINCS algorithm 88
literature-based prediction 136–139
Long Short-Term Memory (LSTM) models 50
low-level DFT method 114

m

machine learning (ML) 111, 131, 150, 165, 185, 208. *see also* deep learning
  algorithms 166
  based methods 59–60, 160–161
  binding affinity prediction 7
  in computational biology 3
  in drug design *9*
  drug discovery 7, 34–35
  in molecular simulation 5
  protein function annotation 4
  in protein function prediction 8
  in protein research 18
  on protein systems 34–39
  quantum mechanics 5–6
machine learning-driven ab initio protein design 165
  advances and applications in 167–170
  graph neural networks for flexible protein design 171–182
  protein sequences *168*
machine learning-guided directed evolution (MLDE) 178
Markov clustering (MCL) 136
masked language modeling 49, 52–53, 189–190
mass spectrometry 219
  mass spectrometry-based proteomics 161
Matthews correlation coefficient (MCC) 193
mCSM method 59
mean absolute error (MAE) 115
membrane proteins, study of 27
metabolic pathways 165
metabolomics 199
metadynamics 24, 26, 59
Meta's ESM2 49–51
Meta's ESMFold 52–54
misfolded proteins 5, 30, 32, 35, 79
MMDesign 178
model interpretability 38
molecular dynamics (MD) simulations 5, 18–19, 24, 82, 100, 113, 134, 168, 185, 188, 207
  protein folding free energy 59
  in protein modeling 8
  traditional folding simulations 79
molecular fractionation 95–96
molecular fragments 84
molecular functions 149, 152
molecular mechanics (MM) 75, 111
  potential energy function *22*
Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) 89
Monte Carlo simulations 33, 82
motif-based approaches 132
multimodal ML 181
multi-omics approaches 151–152
multi-omics data 199
multiple sequence alignments (MSAs) 44, 49
multi-scale modeling 33, 100
mutagenesis 219
MutaPLM 194
mutational patterns 49, 52

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 228 — #6

mutations
  effect on protein structure and function 4
  in neurodegenerative diseases 5
  protein folding free energy change prediction methods 57–70
  protein structure 1
mutual information 132

n

natural language processing (NLP) 137, 152, 188
network-based annotation 159–160, 165, 167
network-based approaches 135–136
network diffusion 136
network propagation for protein function prediction (NPF) method 159, 167
network propagation techniques 166
neural network (NN) model< 66, *80,* 173–174
next-generation sequencing (NGS) technologies 149
N-methylamide (NME) 122
non-bonded interactions 22–23
non-covalent interactions 58
non-machine learning methods 58–59
non-synonymous single-nucleotide polymorphisms (nsSNPs) 57
nuclear magnetic resonance (NMR) spectroscopy 1, 7, 11, 15, 23, 34, 58, 89, 133, 150, 155–156, 170, 185, 207, 211, 217

o

one-hot encoding 38
ontology-based annotation 158–159, *158*
open-access databases 182
OPLS force field 24
"orphan" proteins 151, 209, 212, 215
orthologs 153, 156
overfitting prevention, regularization techniques 99

p

"Pairformer" architecture 176
pairwise interactions 95, 103
pairwise representations 46
PANNZER 158
paralogs 153, 156
PCfun 162
PDZ domain 132
Pearson correlation coefficient (PCC) 59, 66
peptides 35, 169
personalized treatments, protein mutation-based 10
Pfam protein 17, 132, 153–154
physical simulation methods 43
physics-based energy minimization 168
physics-based methods 59, 165
Poisson–Boltzmann (PB) equation 25
polarizable continuum models 33
polarizable force fields 26
position-specific scoring matrices (PSSMs) 69, 132, 185–186
post-translational modifications (PTMs) 35, 176
potential energy surfaces (PESs) 22, 81–82
pre-trained language models 138
pre-training 114
probabilistic frameworks 136
ProGen 185, *186*, 187, 189, 209, 213
propagation-based methods 136
PROSITE 153–154
protein(s) 111, 113, 169, 207, *208*
  amino acid residue interactions *36*
  asphalt molecular structure *25*
  bioinformatics analysis process *8,* 44
  and complexes using AF3 *218*
  docking 133
  dynamics 30
  engineering, advances in 6
  fold databases 16
  folding 1, 7, 23, 24, 26, 30, 33, 43, 57–70, 77–79
  function annotation 4
  misfolding disease 5, 30, 32–33, 35
  modeling, biological datasets for 7
  molecular motion and temperature *32*

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 229 — #7

representation methods *19*
scaffolds 172
sequences 49, 132
simulation 177, *177*
structural levels *210, 211*
structure *2, 10*, 47, *76*
tertiary structure of *20*
three-dimensional (3D) structural model of *16*
time-dependent density functional theory *28*
Protein Data Bank (PDB) 17, 44, 156, 169, 178, 185, 188
protein design
applications in 196–198
future of ESM in 198
protein function annotation, 149, *151*, *152*, 154, 178
high-throughput experimental techniques 161–162
homology-based annotation 156
hybrid methods 162–163
methods and technical framework 152–154
ML-and AI-based methods 160–161
network-based annotation 159–160
ontology-based annotation 158–159, *158*
step-by-step construction process *155*
structural annotation 156–157
workflow of protein function prediction model *157*
protein language models (PLMs)188. *see also* large language models (LLMs)
applications of 192–194
case studies 192–194, *193*
methodological framework for 188–191
protein structure prediction technology *190*
training strategies for 191, *192*
protein–ligand binding affinities 4, 5, 10, 11, 19, 23, 24, 78, 94, 105
protein–ligand docking 7, 21, 31
protein–ligand interactions 26, 29, 33–35, 197
prediction of 113
treatment of 26
protein–molecule interactions 15
protein–protein docking 27
protein–protein interactions (PPIs) 4, 7, 17, 27, 30, 33, 80, 129, *137*, 143–144, *144*, 150, 187, 197, 208
clustering and classification performance of datasets *145*
feature extraction and classification process *130*
future directions 144–146
generating graph networks *133*
graph-based modeling 10
machine learning predicting hot spot regions in *134*
methods and technical framework 131–140
molecular structural features of drugs and proteins *138*
networks 165
results 141–143
protein science 3, 6
applications 214–217
computational 15
protein–small molecule binding method 10
Protein Structure and Interaction Analyzer (PSAIA) 134
protein systems
force field-based calculation on 21–28
ML on 34–39
protein calculations and predictions, strategies for 15–21
quantum mechanical calculation on 28–34
proteomics 155, 199
ProTherm method 68
ProtNote 162
ProtTrans 150, 160, 177, 209, 215
Pseudo-Position Specific Scoring Matrix (PsePSSM) 69
PSI-BLAST software 69, 153
public PPI databases 130

q

quantum algorithms 181
quantum chemistry methods 78
quantum computing 200

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 230 — #8

quantum effects 77–78
quantum-inspired algorithms 178
quantum mechanics (QM) 5–6, 43–44, 81
  calculations 4–8, 11, 19–20, 111
  datasets 77
  descriptions 120
  in protein modeling 18
  in protein simulations 11
  on protein systems 28–34, *30*
quantum mechanics/molecular mechanics (QM/MM) 27–30, 75–76, 83, 112
quantum properties 77, 78

r

Random Forest models 140
random walk-based algorithms 136
recurrent neural networks (RNNs) 6, 35, 132, 140, 175, 189
reinforcement learning (RL) 37, 152, 169, 209
replica exchange molecular dynamics (REMD) 26, 59, 88, 114, 122
residue gas state 47
resource-intensive screening 89
ribosome profiling 219
root-mean-square deviation (RMSD) 88
root-mean-square error (RMSE) 115
root-mean-square fluctuation (RMSF) 88
RoseTTAFold 113, 134, 176

s

SAAFEC method 59
Saccharomyces cerevisiae 132
salt bridges 23
Schrödinger equation 28–29, 75, 76
SciBERT process 138
SciSpacy 137
self-distillation 48
self-supervised learning 178
self-supervision 49
semi-empirical methods 29
sequence-based methods 131–132, 141
sequence-based PPI prediction methods 17–18
sequence optimization 197
sequence similarity 153
sequence–structure–function paradigm 5, 17
sequence-to-sequence prediction 190
SH2 domain 132
signal transduction 30, 32, 80
SITEHOUND identify interaction interfaces 134
small molecule binding (SMB) 192
small molecules
  comparison with proteins in modeling 78–79
  library 89
SMART 153–154
solvent effects 33
spectroscopy 58, 207
standard force fields 26
statistical modeling 207
STRING database 130, 137–138, 154, 155, 159, 166
structural annotation 156–157
  databases 150
  methods 209
structural biology techniques 155
Structural Classification of Proteins (SCOP) 150, 153, 157
structure-based methods 133–135, 141
structure prediction model 47
supervised learning 37, 212
  on labelled data 48
support vector machines (SVMs) 35
Support Vector Regression (SVR) 66
surface plasmon resonance (SPR) 218
synthetic biology, protein design for 198
synthetic biomaterials 169
synthetic data generation 169

t

Tasks Assessing Protein Embeddings (TAPE) 190
t-distributed stochastic neighbor embedding (t-SNE) method 50, *50, 51,* 64
Tensor Processing Unit (TPUs) 189
text mining 136–139
therapeutic applications of proteins 197

"bind_PrintPDF" — 2025/9/15 — 19:13 — page 231 — #9

therapeutic interventions, protein-based 7
therapeutic strategies
  mutation-based 4
  protein interaction targeting 80
threading (fold recognition) 16–17, 134–135
three-dimensional (3D) structures 113, 130, 165, 185, 207
torsional rotations 22
traditional classical force fields 112
transcriptomics 199
transfer learning (TL) 111–112, 169
  classical methods to TL in protein FQM calculations 112–113
  deep learning protocol for FQM calculations 120–127, *123*
  inductive transfer learning force field 113–120
  in protein modeling 11, 21, 37
  transferability assessment and statistics and distribution of datasets *125*
transformers 132, 189
  model 49, 138, 195, 212
  transformer-based architectures 131, 209
transition state modeling 113
"triangle multiplicative update" model 46
two-body molecular fractionation method 95–98, 101–103, 106

u

Uniclust 30, 48
UniParc database 49
UniProt database 60, 155, 178, 186, 188–189
unlabelled data 48, 49
unsupervised learning 49–50, 52–53, 212
user-friendly software 182

v

van der Waals force interactions 23, 58
Variational Autoencoders (VAEs) 168, 174, *174*, 178, 209
variational graph autoencoders (VGAEs) 171
virtual screening process 79

w

water–protein interactions 25
weight regularization 122

x

XGBoost (XGB) 60, 65, *65,* 66, **67**, *67, 68,* 69
X-ray crystallography 1, 7, 11, 15, 23, 34, 89, 150, 155–156, 170, 185, 207, 211, 217, 219

y

yeast two-hybrid (Y2H) assays 129, 208, 212

z

ZDOCK 133
ZINC database 89