

Index

A

- Abbreviations, 2
 - Activation maximization, 125, 128–129
 - AdabmDCA, 116
 - Aleatoric uncertainty, 95
 - Algorithms
 - carbon footprint of, 154–156
 - CASP15 servers use of prediction algorithms, 62–64
 - design algorithm considerations, 98–101
 - estimation of distribution algorithm, 99
 - Allosteric regulation, engineering, 45–46
 - ALMs. *See* Antibody language models
 - AlphaFold
 - backbone conditional protein sequence design, 145–146
 - carbon footprint of algorithms, 155
 - FAPE loss, 133
 - impact of, 109
 - as leap forward, 151, 155
 - protein design using, 125–137
 - as protein structure foundation model, 135–137
 - recycling steps, increasing number of, 145
 - structural metrics predicted by, 144–145
 - AlphaFold2
 - antibody language models (ALMs) compared, 24–25
 - as breakthrough, 4–6
 - CASP challenges, 53–55
 - EvoFormer, 7–8, 25
 - experimental and structural biology changed by, 4–5
 - multimer modeling, 61
 - multiple sequence alignment (MSA) use, 4–11, 54
 - new solutions boosting power of, 5–6
 - pLMs compared, 9
 - resolution of structure space, 5
 - AlphaFoldDB database, 5–6, 11
 - AlphaFold-Multimer, 5
 - AlphaSeq, 26
 - ANNs (artificial feedforward neural networks), 2
 - AntiBERTa, 23, 29
 - AntiBERTy, 24, 27, 29
 - Antibody engineering
 - binding affinity optimization, 26
 - multiparameter optimization, 27
 - overview, 25–27
 - safety, 26–27
 - Antibody-GAN, 115
 - Antibody humanization, 26–27
 - Antibody language models (ALMs)
 - antibody engineering, 25–27
 - binding affinity optimization, 26
 - challenges and opportunities, 27–29
 - evolution of, 20–24
 - generative, 21, 24
 - large-scale data sets for machine learning, creating, 28
 - pretraining, 22–24
 - structure prediction driven by, 24–25
 - technical standardization, 28–29
 - transformer-based, 21
 - Antibody sequence representation
 - evolution of models, 21–24
 - natural language processing, 19–20
 - Artificial feedforward neural networks (ANNs), 2
 - Assay-labeled data, 90–91
 - Attention mechanism, 109
 - ATUE, 28
 - Autoregressive models, 110–112, 114, 116, 146
- ## B
- Backbone conditional protein sequence design
 - evaluating performance methods, 143–145
 - graphic models, 143
 - large language models, 146–147
 - missing background information, 146–147
 - overview, 142–143
 - review, 141–147
 - symmetries and multistate design, 146
 - training data and overfitting, 145–146
 - Bayesian optimization, 98–99
 - B-cell receptors (BCRs), 17–19, 23, 28
 - B cells, 17–18
 - BFD (Big Fantastic Database), 4, 8, 10, 55, 57, 59
 - BFD/Mgnify, 55
 - Bias
 - in backbone conditional protein sequence design, 144
 - incorporating informative inductive, 93–94
 - Bidirectional encoder representations from transformers (BERTs), 21–23, 26, 109–110
 - Big Fantastic Database (BFD), 4, 8, 10, 55, 57, 59
 - BLAST, 54
 - BmDCA, 116
 - Bolt-LMM, 157
- ## C
- Carbon footprint
 - of algorithms, 154–156
 - calculating, 152
 - of data storage, 153
 - energy usage, 153

Index

- Carbon footprint (*Continued*)
 - estimating in practice, 154
 - of experimental work, 154
 - molecular simulations, 155
 - protein–protein interactions, 155
 - CARP (convolutional autoencoding representations of proteins), 113, 147
 - CASP (critical assessment of protein structure prediction), 4, 25, 53–55, 57, 65
 - CASP15 servers
 - petabase-scale homology search, 55–57, 59, 65
 - size and composition of reference databases used by, 56
 - strategy selection and comparison with, 65
 - use of homology algorithms and databases, 55
 - use of prediction algorithms and strategies, 62–64
 - Causal language modeling (CLM), 20, 22
 - Causal models, 20, 22, 89, 93, 95
 - CDRs (complementarity determining regions), 17, 20–21, 24, 26, 28
 - ChatGPT, 109
 - Chroma, 133
 - CLM (casual language modeling), 22, 24
 - Closed-form density, 97
 - CNNs (convolutional neural networks), 112–113, 119
 - CodeCarbon, 154
 - Coevolution-based computational protein design, 35–48
 - ColabFold, 5–6, 9, 55
 - ColabFoldDB, 55, 57
 - ColabFold-predict, 53, 55, 57, 59, 61, 65–66
 - ColabFold-search, 53, 55, 57
 - Comparative modeling, 3, 5
 - Complementarity determining regions (CDRs), 17, 20, 24, 26, 28
 - Computational protein design, coevolution-based, 35–48
 - Computations
 - coevolution-based computational protein design, 35–48
 - environmental impacts of, 152–154
 - Conditional language models, 111–116
 - Conditional sequence design. *See* Backbone conditional protein sequence design
 - Conditional transformer language (CTRL), 111–112
 - Confusion matrix, 144
 - Convolutional autoencoding representations of proteins (CARP), 113, 147
 - Convolutional neural networks (CNNs), 20–21, 112–113, 119
 - Covariate shift, 93
 - Critical assessment of protein structure prediction (CASP), 4, 25, 53–55, 57, 65
- ## D
- DARK, 111
 - Dark proteome, 9
 - Data centers, carbon footprint of, 152–154
 - DCA. *See* Direct coupling analysis
 - DDPMs (denoising diffusion probabilistic models), 132–134
 - Decoder-only transformers, 110, 112
 - Deep learning
 - paradigm shift in protein design, 108–109
 - unsupervised, 69–70, 72, 77, 79–83, 116
 - Deep mutational scanning, 36, 45, 71–72, 80, 90
 - DeepSequence, 47, 70, 115, 119–120
 - Denoising diffusion probabilistic models (DDPMs), 132–134
 - De novo protein design methods, schematic of, 127
 - Density ratio estimation, 90, 97
 - Design algorithm considerations
 - Bayesian optimization, 98–99
 - estimation of distribution algorithm, 99
 - hyperparameter selection, 99–101
 - for novel property values, 98–101
 - sequence diversity, 100–101
 - Design-induced distribution shift, accounting for, 93
 - Designing proteins with novel property values
 - challenges, 87–89
 - design algorithm consideration, 98–101
 - future of, 101–102
 - learning a trustworthy model, 89–94
 - quantifying uncertainty, 94–98
 - Diffusion models, 132–134
 - Direct coupling analysis (DCA)
 - alignment composition, 37–39
 - applications, 46
 - comparisons between contact prediction and structural data, 45
 - de novo sequence design and testing, 46–47
 - engineering allosteric regulation, 46
 - evaluation of, 42–47
 - general overview, 36–37
 - mathematical overview, 38, 40–42
 - open problems and future directions, 47–48
 - prediction of mutational fitness effects, 45
 - Directed evolution
 - machine learning-assisted (MLDE), 117–119
 - novel property values, 87–88
 - techniques and protein design, 116–120
 - traditional (TDE), 117–118
 - Distograms, 4
 - Distribution shift, 89, 93, 95–98
- ## E
- Electricity consumption, 153
 - Embeddings, 7–10, 22–25
 - Emergent behavior, 102
 - Encoder-only protein language models, 110, 112
 - Environmental impacts, of machine learning applications, 151–157
 - EquiFold, 25
 - ESM (evolutionary scale modeling)
 - ESM-2, 80, 110
 - ESM-1b, 27, 119–120, 147
 - ESMFold, 5–6, 9, 11, 126, 155–157
 - ESM-IF1, 147
 - variant effect identification, 76–77, 80, 83–84

- Estimation of distribution algorithm, 99
- EV mutation, 47
- EvoFormer, 7–8, 25
- Evolutionary data, uncertainty quantification for models of, 97–98
- Evolutionary fitness, variant effect and, 71–74, 78
- Evolutionary information, 2, 4, 8–11
- Evolutionary scale modeling. *See* ESM
- E-waste, 152–153

- F**
- FAPE (frame-aligned point error), 129, 133–134
- FBGAN, 115
- Feedback covariate shift, 93, 96
- Fixed backbone sequence design, 128, 131
- FoldingDiff, 133
- Foldseek, 5–6, 11
- Forward folding, 128
- Frame-aligned point error (FAPE), 129, 133–134
- FrameDiff, 133–134
- Free hallucination, 128–129
- Free modeling, 54

- G**
- Gaussian process regression models (GP-UCB), 98–99
- Generative adversarial networks (GANs), 20–21, 109, 113–115, 131–132
- Generative models
 - conditional language, 111–116
 - diffusion, 132–134
 - review, 107–120
 - traditional approaches compared, 107–109
 - transformer-based language models, 109–112
 - unconditional language, 110–111
- Generative pretrained transformer (GPT)
 - GPT-2, 21–22, 111
 - GPT-4, 109
 - transformer-based language models, 109–111
- Generative protein sequence models (GPSMs), 43, 47–48
- Genie, 133
- Global minimum energy conformation (GMEC), 107–108
- Global warming potential (GWP), 152
- GLUE, 28
- GPSMs (generative protein sequence models), 43, 47–48
- GPT. *See* Generative pretrained transformer
- GP-UCB (Gaussian process regression models), 98–99
- Green Algorithms, 154
- GREENER principles, 157
- Greenhouse gas (GHG) emissions, 152, 154–155, 157
- GWP (global warming potential), 152

- H**
- Hallucination, 128–136
- HHblits, 54–55, 57, 59
- HHpred, 54
- HHsearch, 54, 61
- Hidden Markov model (HMM), 21, 54, 120
- Homologous sequence data, as model training data, 91
- Homology search, petabase-scale, 53–66
- Humanization, antibody, 26–27
- Hyperparameter, 28–29, 91, 97, 99–100

- I**
- IgFold, 24–25
- IgLM, 26–27
- ImageNet, 28
- IMG/M database, 55
- ImmuneBuilder, 25
- Inverse folding, 107, 113, 128
- iReceptor database, 19

- L**
- Language models
 - for antibody comprehension, 20–21
 - architectural types, 110
 - conditional, 111–116
 - population frequency, recognition of, 73–75
 - pretraining and fine-tuning regime for, 21
 - protein (pLMs) (*see* Protein language models)
 - summary of released, 112
 - transformer-based, 109–113, 131
 - unconditional, 110–111
- Large language models (LLMs)
 - future for novel protein design, 101–102
 - missing backbone information, 147
- Long short-term memory (LSTM) networks, 112, 114

- M**
- Machine learning applications, environmental impacts of, 151–157
- Machine learning-assisted directed evolution (MLDE), 117–119
- Markov chain Monte Carlo (MCMC), 110, 141
- Masked inverse folding (MIF), 113
- Masked language modeling (MLM), 22–24, 113, 146–147
- Metagenomics, 55
- MetaRNN, 72–79, 81, 83
- Mi3, 116
- MIF (masked inverse folding), 113
- MLDE (machine learning-assisted directed evolution), 117–119
- MLM (masked language modeling), 22–24, 113, 146–147
- MMseqs2, 6, 9–10, 55, 57, 59, 146
- MODELLER, 54
- Molecular simulations, carbon footprint of, 155
- MSA-Transformer, 7
- MSAVAE, 115
- Multimer modeling, 61–65

Index

- Multiple sequence alignment (MSA)
 - AlphaFold2 use, 4–11
 - avoidance in antibody language models (ALMs), 24–25
 - enrichment using the Sequence Read Archive (SRA), 58
 - evolutionary information derived from, 2
 - generative models based on, 114–116
 - petabase-scale homology search, 53–55, 57–66
 - PSI-BLAST, 54
 - strategies for statistical analysis, 36–42
 - transformer, 110
- Multistate backbone design, 146
- Mutational fitness effects, prediction of, 45
- MVP, 73, 78, 82

- N**
- Native sequence recovery (NSR), 143–144
- Natural language processing (NLP)
 - antibody sequence representation, 19–21
 - architectural types of language models, 110
 - protein language models (pLMs) spawned from, 6–7
 - self-attention heads, 28
- Next-generation sequencing (NGS), 18, 45
- Nonprotein molecules, protein design in context of, 134–135
- Novel conditions, learning a trustworthy model for
 - design-induced distribution shift, accounting for, 93
 - inductive biases, incorporating informative, 93–94
- Novelty
 - challenges in finding, 87–89
 - radical, 88
 - spectrum, 88
 - types of novel conditions, 87–88
- NSR (native sequence recovery), 143–144

- O**
- Observed Antibody Space (OAS), 19

- P**
- Pan-protein data, as model training, 91–92
- Pathogenicity, variant effect and, 70–74, 78, 80, 82–84
- PepVAE, 115
- Petabase-scale homology search
 - effect of homologs on structure prediction, 59
 - homolog search and MSA construction, 57–59
 - introduction, 53–55
 - leveraging templates, 59–60
 - multimer modeling, 61–65
 - review, 53–66
 - tuning parameters, 59–65
- PhastCons, 73, 75–78, 80–83
- PHD, 2, 4
- pLMs. *See* Protein language models
- PMD (Protein Mutant Database), 70, 78–80
- Potts model
 - closed-form density, 97
 - designing proteins with novel property values, 90–91, 97–98
 - direct coupling analysis (DCA), 35–37, 42–43
 - as generative model, 115–116
 - joint amino acid distribution, capturing, 142
 - pairwise statistics, 145
 - Potts Hamiltonian models, 37, 115–116
- Prediction error, 90
- ProGen2, 23, 111
- ProGen2-OAS, 23, 27
- ProtDiff, 132–133
- Protein alignment methods, retirement of traditional, 10–11
- Protein design, in context of nonprotein molecules, 134–135
- ProteinGAN, 113
- ProteinGenerator, 134–135
- Protein language models (pLMs)
 - antibody language models (ALMs) compared, 21
 - embeddings, 7–10
 - paradigm shifts triggered by, 6–11
 - protein-specific predictions, 9
 - transformers for building, 23
 - variant effect identification, 72, 77–78, 80
- ProteinMPNN, 129, 131–132, 134, 136, 144–145
- Protein Mutant Database (PMD), 70, 78–80
- Protein–protein interactions
 - carbon footprint of predicting, 155
 - permanent, 4–5
- ProteinVAE, 115
- ProteoGAN, 115
- ProtGPT2, 111
- Protpardelle, 133
- ProtT5, 10, 111
- ProtTrans
 - ProtTransT5, 75, 78–80
 - variant effect prediction, 75–77, 80–81, 83
- PSI-BLAST, 54

- R**
- REGENIE, 157
- REVEL, 70, 72, 75–76, 78–81, 83
- RFAA (RF all-atom), 135–136
- RF_{diffusion}, 133–136
- RF_{joint}, 132–135
- RFNA (RF nucleic acid), 135
- RITA, 111
- Rosetta energy minimization, 141
- Rosetta FastDesign protocol, 141
- RoseTTAFold
 - ESMFold compared, 9
 - inpainting missing information, 147
 - protein design using, 125–137
 - as protein structure foundation model, 135–137

- S**
- SabDab, 28
- Sapiens ALM, 26–27

- SCA. *See* Statistical coupling analysis
 - Secondary structure prediction, rise in, 2–3
 - Seq2seq models, 109
 - Sequence design methods, evaluating performance of
 - bias and confusion matrix, 144
 - native sequence recovery (NSR), 143–144
 - sequence diversity and pairwise statistics, 145
 - sequence–structure compatibility, 144–145
 - Sequence diversity, 24, 38, 43, 100–101, 145
 - Sequence Read Archive (SRA)
 - multiple sequence alignment enrichment using, 58
 - petabase-scale homology search of, 53, 55, 57–66
 - Sequence–structure compatibility, 144–145
 - Sequencing
 - B-cell receptors, 18–19
 - decreasing cost of, 101
 - deep mutational scanning, 90
 - metagenomics, 55
 - next-generation, 18, 45
 - single-cell, 18–19
 - SeqUNet, 72–73, 76–77, 82
 - Serratus, 57, 59, 66
 - SEWING, 141
 - Sickle cell hemoglobin allele, 71
 - SIFT, 72–73, 76–77, 79–81
 - Single-nucleotide variant (SNV), 69–70, 84
 - Single-sequence structure prediction, 126
 - Singular value decomposition (SVD), 38, 42
 - SNAP, 70
 - SRA. *See* Sequence Read Archive
 - Statistical coupling analysis (SCA), 35–48
 - alignment composition, 37–39
 - applications, 46
 - de novo sequence design and testing, 46–47
 - engineering allosteric regulation, 45–46
 - evaluation of, 42–47
 - general overview, 36–37
 - mathematical overview, 38, 40–42
 - open problems and future directions, 47–48
 - Statistical models of protein sequence
 - assessing model fit and generative capacity, 43
 - coevolution-based computational protein design, 35–48
 - direct coupling analysis (DCA), 35–38, 40–42, 45–48
 - evaluation of, 42–47
 - experimental testing of computational protein design, 43–47
 - generative protein sequence models (GPSMs), 43, 47–48
 - higher-order covariance, 43
 - open problems and future directions, 47–48
 - sequence diversity, 43
 - statistical coupling analysis (SCA), 35–48
 - strategies, 36–42
 - Structure-based protein design
 - joint, 136
 - steps in, 128
 - Structure prediction
 - carbon footprint of algorithms, 155–156
 - as design filter, 126, 128
 - petabase-scale homology search for, 53–66
 - as pretraining, 131–132
 - Structure-prediction networks, protein design using, 125–137
 - SVD (singular value decomposition), 38, 42
 - Symmetric sequences, 146
- ## T
- T-Coffee, 10
 - Temperature, 145
 - Template-based modeling (TBM), 54
 - Three-state per-residue accuracy, 2–3
 - Traditional directed evolution (TDE), 117–118
 - Training data
 - assay-labeled data, 90–91
 - homologous sequence data, 91
 - pan-protein data, 91–93
 - trade-off between quality and quantity, 89–93
 - Transfer learning, 6–9
 - Transformer-based language models, 109–113, 131
 - Transformer neural networks, 21
 - Transformers, 8, 21–23, 25–29, 109–112, 117
 - TRUST4, 18
- ## U
- Uncertainty, aleatoric, 95
 - Uncertainty quantification
 - Bayesian, 94–96
 - frequentist, 95–97
 - for models of evolutionary data, 97–98
 - usefulness of, 94
 - Unconditional structure generation, 128
 - Unconditioned generative language models, 110–111
 - UniProt, 4, 8
 - UniRep, 76, 112, 119–120
 - Unsupervised methods
 - DeepSequence as, 119
 - variant effect prediction, 69–70, 72, 75, 77–84
- ## V
- Variant effect
 - annotation with tools of today, 72–73
 - conservation, 77–78
 - fitness, 71–74
 - functional, 70–73, 78, 80, 82–83
 - impact of type on method development, 71–72
 - kinds of effects, 70–71
 - pathogenic, 70–74, 78, 80, 82–84
 - Variant effect prediction
 - capture of different signals by methods, 80
 - case for machine learning, 70
 - conservation as orthogonal to frequency and recognized by all predictors, 77–78

Index

Variant effect prediction (*Continued*)
correlation across methods, 75–84
correlation of variant predictor scores, 76
effects as correlated for, 83–84
functional effect as combination of factors recognizable
 by unsupervised methods, 78, 80
pathogenicity signals captured by, 80, 82–83
population frequency, 73–75
review, 69–84
Variant learning, 69–84
Variational autoencoders (VAEs), 21, 115, 131–132
Vector quantised-variational autoencoder
 (VQ-VAE), 6

W

Word2vec, 20

X

xTrimoPGLM-Ab, 23–24, 29
xTrimoPGLM-AbFold, 24–25

Z

Zero-shot, 91–92, 111
ZymCTRL, 112