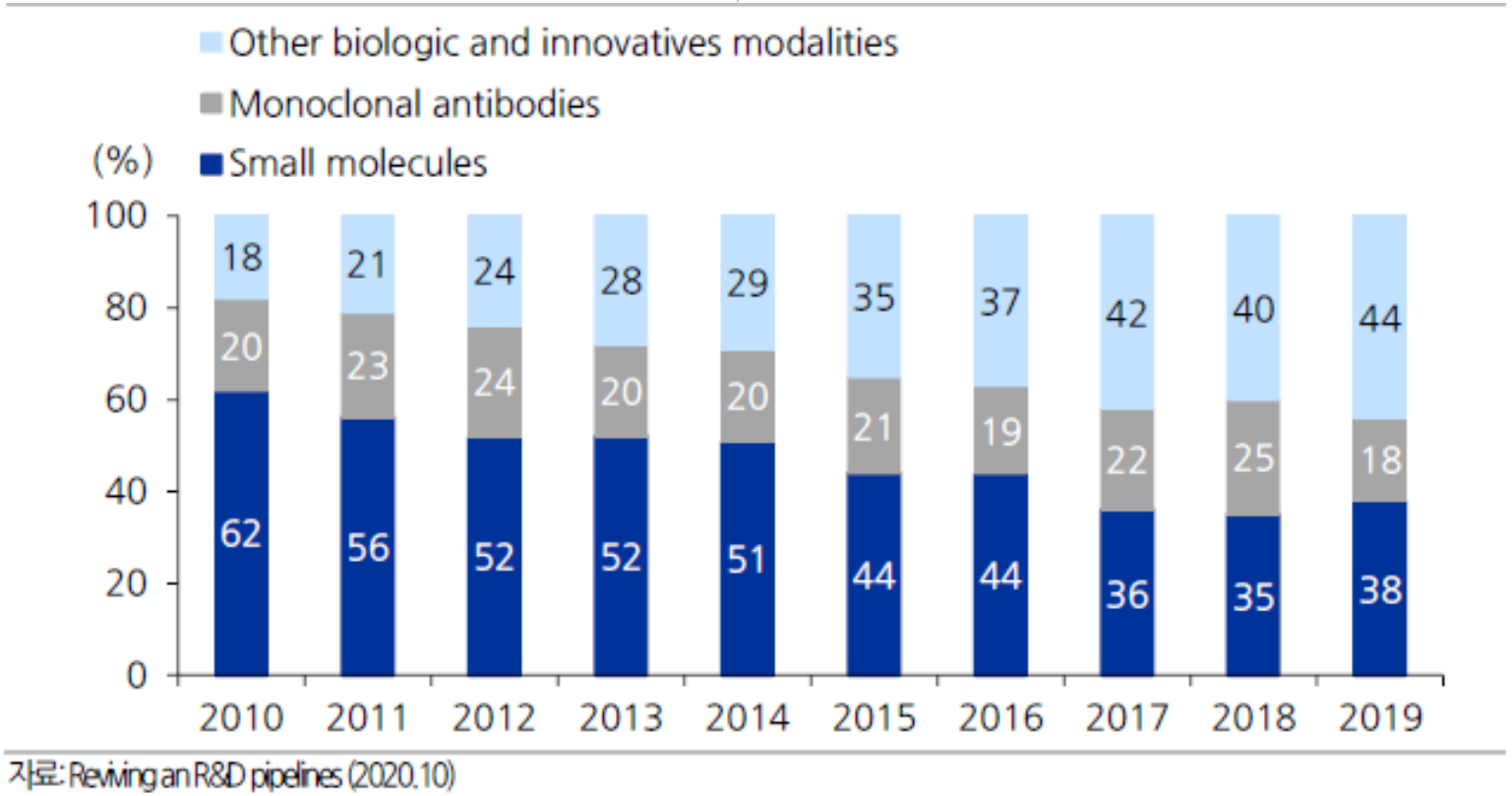
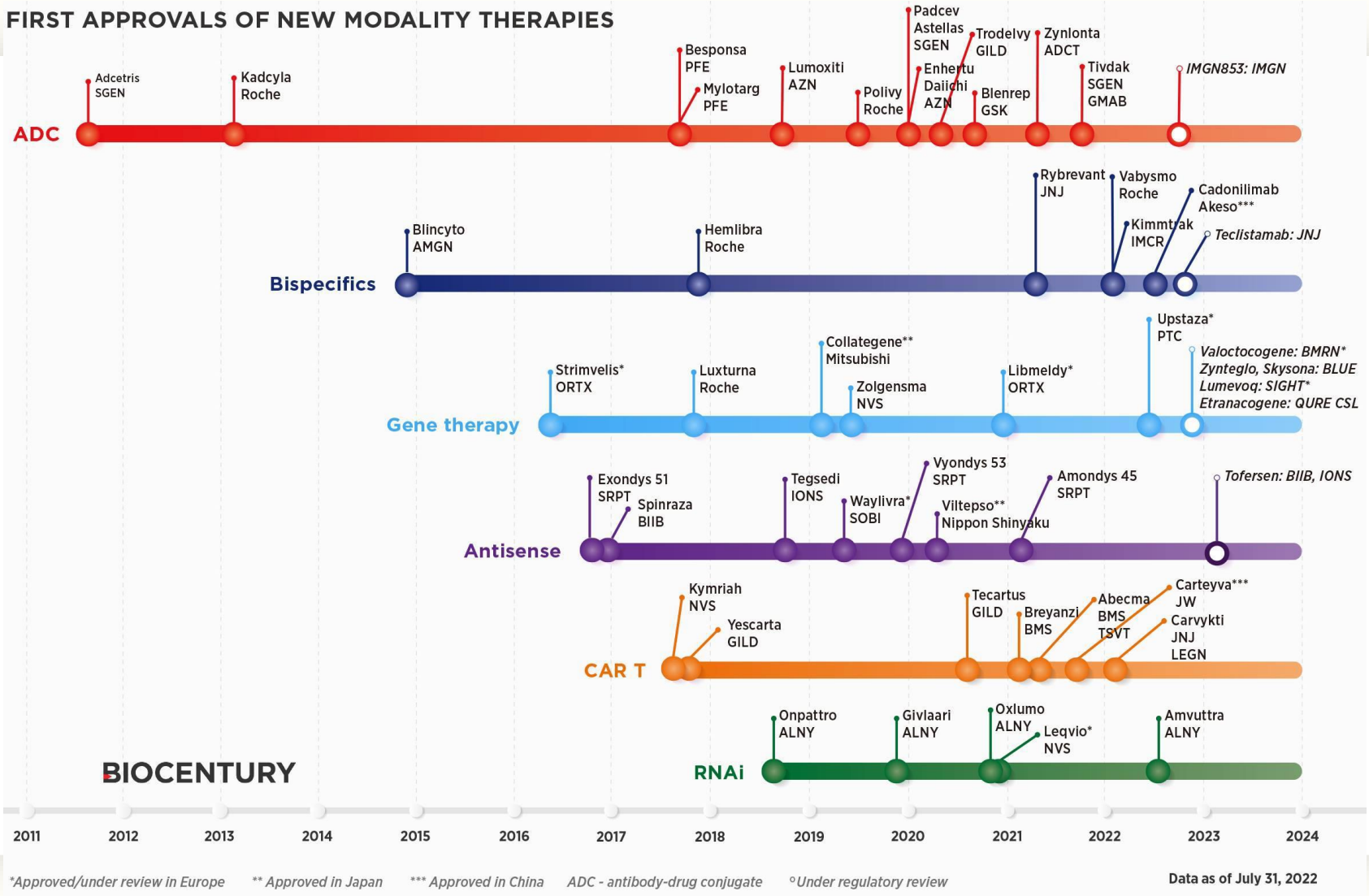
- 07

- **혁신 신약 모달리티의 이해**



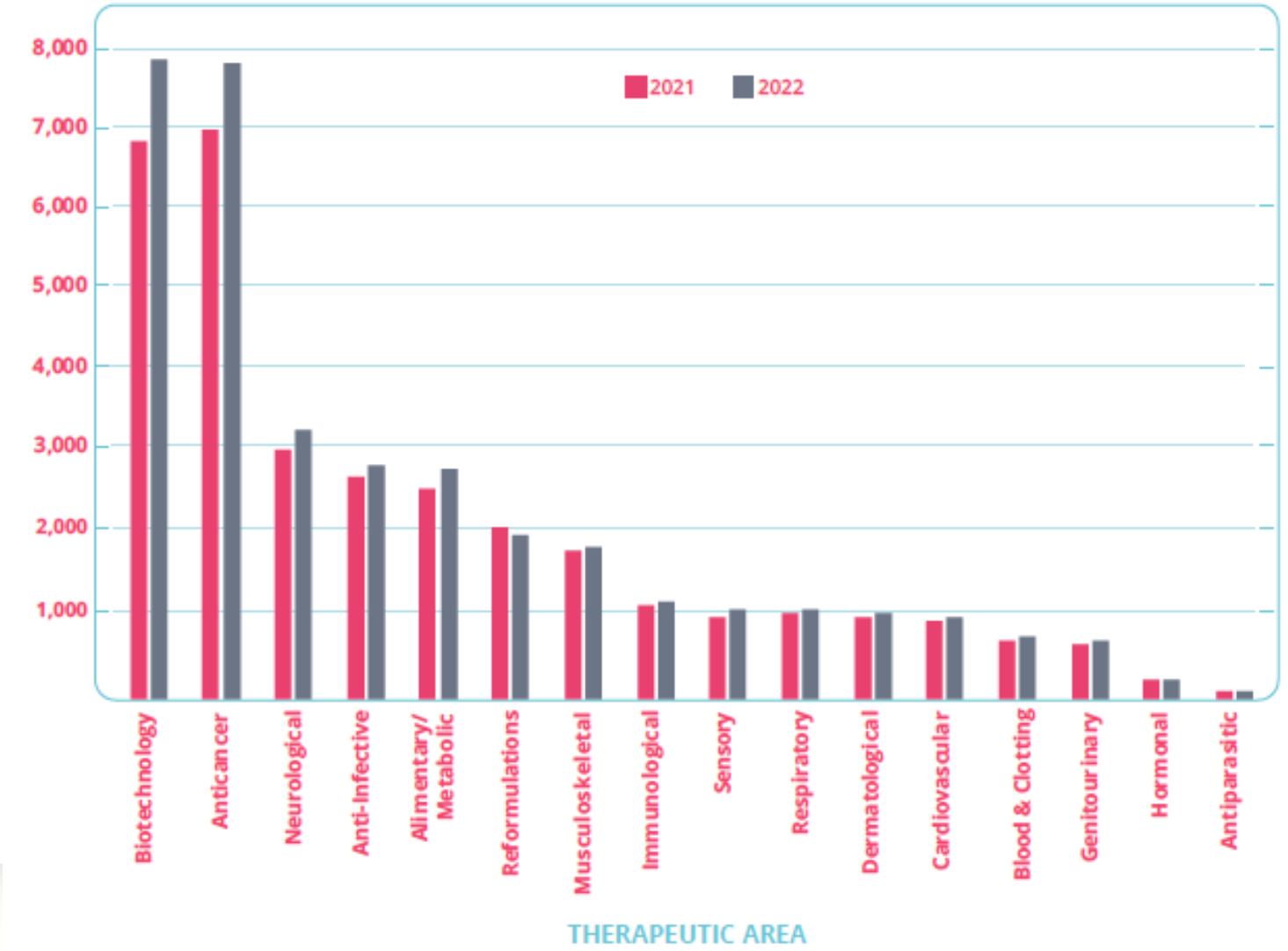
자료: Reviving an R&D pipelines (2020.10)

* Source: xxxx, 2022.0x.xx



* Source: xxxx, 2022.0x.xx

The R&D pipeline by therapy group, 2021 and 2022



* Source: xxxx, 2022.0x.xx

Computational and Structural Biotechnology Journal 20 (2022) 5316–5341

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Review

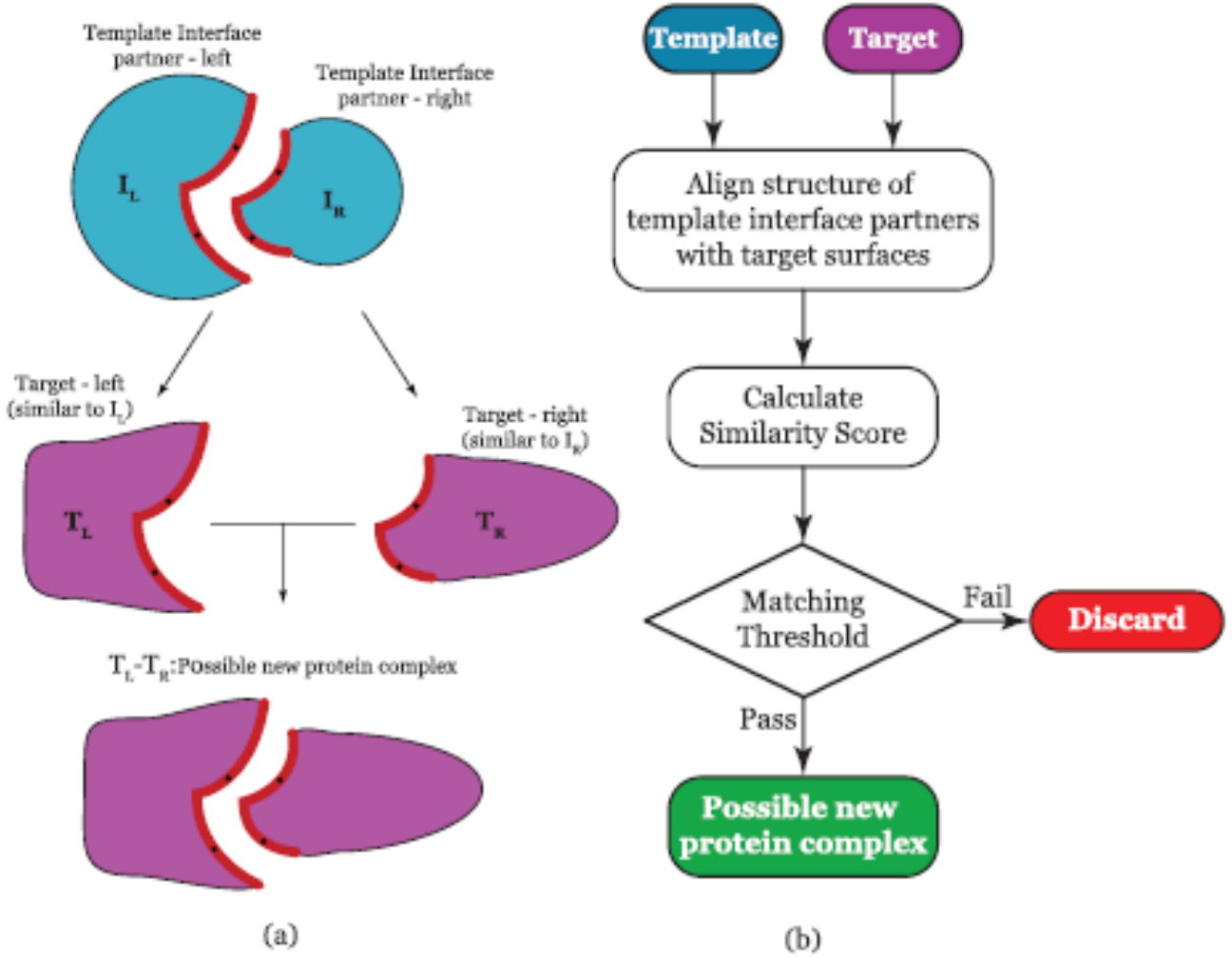
Protein–protein interaction prediction with deep learning: A comprehensive review

Farzan Soleymani [a], Eric Paquet [b,*], Herna Viktor [c], Wojtek Michalowski [d], Davide Spinello [a]

[a] Department of Mechanical Engineering, University of Ottawa, Ottawa, ON, Canada
[b] National Research Council, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
[c] School of Electrical Engineering and Computer Science, University of Ottawa, ON, Canada
[d] Telfer School of Management, University of Ottawa, ON, K1N 6N5, Canada

Table 4
Databases for PPI prediction

Type	Database	Description	Last update	URL
Protein–Protein Interactions	STRING [109]	Functional associations between protein pairs, which contains 67,592,464 proteins from 14094 organisms; 20,052,394,042 interactions.	2021	https://string-db.org/
	IntAct [116]	Contains manually curated datasets (topical), interactomes (for 16 different species) and annotations of experimental evidence.	2021	https://www.ebi.ac.uk/intact/home
	Biogrid [111]	Contains 2,467,140 protein and genetic interactions, 29,417 chemical interactions and 1,128,339 PTMs from major model organism species.	2020	http://www.thebiogrid.org/
	DIP [112]	Experimentally determined PPI database including biological information of proteins, PPIs and experimental techniques for identifying interactions.	2020	https://dip.doe-mbi.ucla.edu/dip/Main.cgi
	Negatome 2.0 [108]	Contains 21,795 interactions, with scores of zero and one, using text mining from literature and analysing protein complexes from PDB.	2014	http://mips.helmholtz-muenchen.de/proj/ppi/negatome/
	MINT [114]	Experimentally curated PPI database that includes approximately 117001 PPIs from 607 different species.	2012	https://mint.bio.uniroma2.it/
	HPRD [115]	Consists of 41,327 PPIs, 93,710 PTMs, 22,490 Subcellular Localizations and 112,158 Protein Expressions.	2010	http://www.hprd.org
	BIND [113]	PPIs collected from of humans, yeasts, nematodes, etc.	2005	http://download.baderlab.org/BINDTranslation
Protein sequences	UniProt [106]	A collection of protein sequence and functional information, including UniProtKB, UniParc, UniRef and Proteomes. UniProtKB contains 567,483 reviewed (Swiss-Prot)—manually annotated, and 231,354,261 unreviewed (TrEMBL)—computationally analysed, protein sequences.	2020	http://www.uniprot.org
	SWISS-MODEL [107]	A web-based integrated service providing information for protein structure homology modelling. The repository contains 2,217,470 models from SWISS-MODEL for UniProtKB targets, as well as 180,107 structures from PDB with mapping to UniProtKB.	2020	https://swissmodel.expasy.org/
	PIR [119]	Integrated protein resources, including protein sequences and high-quality annotations by integrating more than 90 biological databases.	2022	http://pir.georgetown.edu/
Higher-level structures	RCSB PDB [110]	Information about the 3-D structure of proteins, nucleic acids, and complex assemblies. 191144 structures, 57349 human sequence structures, and 14406 nucleic acid-containing Structures	2021	https://www.rcsb.org/
	SCOP [122]	Classification of known proteins and a comprehensive description of the structural and evolutionary relationships between them. As of 2022-05-30, this dataset contains 72,448 non-redundant domains, representing 858,316 protein structures.	2022	http://scop.mrc-lmb.cam.ac.uk/scop
Genomic information	CGD [124]	A resource for genomic sequence data, genes and protein information for *Candida albicans* and related species.	2022	http://www.candidagenome.org/


Template Interface partner - left
Template Interface partner - right
I_L
I_R
Target - left (similar to I_L)
Target - right (similar to I_R)
T_L
T_R
T_L-T_R:Possible new protein complex
(a)
Template
Target
Align structure of template interface partners with target surfaces
Calculate Similarity Score
Matching Threshold
Fail
Discard
Pass
Possible new protein complex
(b)

Framework	Description	Advantage	Disadvantage
GCN-based [2017] [247]	This study proposed a pairwise classification architecture in which one or more graph convolution layers process the neighbourhood of a residue in each protein. Then, the representation of two residues is paired and passed through a dense layer for classification. This study analysed several GCN-based methods, concluding that neighbourhood-based convolution methods outperform diffusion-based convolution and SVM-based methods.	The proposed convolution operators and obtained features may be helpful for other applications, including protein function, catalytic and other functional residues, and protein interactions with DNA and RNA.	The accuracy of this approach is examined based on a limited number of labelled training examples.
IntPred [2018] [240]	This method uses a random forest to predict protein–protein interface sites at both the surface patch and residue levels.	The performance of a binary classifier can be evaluated using different measurements, such as the Matthews' correlation coefficient (MCC), sensitivity, precision, and specificity [240]. IntPred outperformed the methods ProMate [189], PIER [249], PINUP [250], and meta-PPISP [251], but not SPPIDER [252], based on MCC.	The performance of this method depends on the application. For instance, IntPred was better suited to cases in which false positives are less well tolerated than false negatives.
Graph-based generative model [2019] [177]	This method uses a graph transformer model for designing protein sequences given graph representations of 3-D protein structures, leveraging the spatial locality of dependencies in molecular structures.	This method uses a self-attention mechanism to capture higher-order, interaction-based dependencies between sequence and structure. The graph-based model offers computational efficiency due to the representation of long-range sequence dependencies by short-range sequence dependencies in 3-D space [253–255]. Additionally, they achieved linear computational scaling concerning the sequence length and representational flexibility for coarse and fine-grained structure descriptions.	The evaluation dataset only contained chains up to a length of 500, limiting the applicability of this approach.
Struct2Graph [2020] [248]	In this method, graph embeddings of each protein are obtained using an assigned GCN. Next, relevant geometric features associated with query protein pairs are extracted using a mutual attention network. Finally, a feedforward neural network performs a binary classification between interacting and noninteracting pairs.	Struct2Graph only uses 3-D structural information to predict the PPI. They have reported state-of-the-art performance on both balanced and unbalanced datasets.	Limited availability of 3-D structural information may restrict the applicability of this method.
LSTM-based [2020] [187]	The proposed method integrates the 3-D structure and sequence-based information of proteins to predict PPIs. The 3-D coordinate information, hydropathy index, isoelectric point, and amino acid charges of each protein are fed into a pre-trained ResNet50 model to extract features from these attributes. A stacked autoencoder obtains the compact form of encoded proteins using autocovariance and conjoint triad. The structural features from ResNet50 are passed through LSTM and concatenated with features from the stacked autoencoder. The merged features are then fed into the classifier to predict protein pair labels.	This method performs well despite being trained on a low number of instances.	Limited availability of 3-D structural information may restrict the applicability of this method. Additionally, LSTM models are computationally demanding and slow.

Summary of advantages and disadvantages of sequence-based Deep Learning methods regarding PPI prediction.

Framework	Description	Advantage	Disadvantage
SVM-conjoint triad [2007] [286]	Each protein sequence was represented in this study by a vector of amino acid features. The model was developed based on a support vector machine (SVM) integrated with a kernel function and a conjoint triad feature for describing amino acids. This method mapped different types of PPI networks using only sequence information, which could be applied to explore networks for any newly discovered protein with unknown biological relationships. They suggested that methods without local environments for amino acids are often unreliable, so a conjoint triad method was used.	The 20 standard amino acids were clustered into several classes based on their dipoles and side chain volumes to achieve dimensionality reduction of the vector space. This method might predict PPI networks created by pairwise PPIs.	The limited available information on protein pairs restricts the applicability of this method. Additionally, it mainly considers the properties of two nearby amino acids, overlooking long-range interactions.
SVM-autocovariance [2008] [288]	This method combined a new feature representation using autocovariance (AC) and a support vector machine (SVM). AC considers the interactions between more distant amino acids in the protein sequence, specifically long-range interactions. This is an improvement over the method proposed in [286]. This model was evaluated using an independent dataset of 11,474 yeast PPIs.	The conjoint triad (CT) method only considered the attributes of an amino acid and its two neighbouring amino acids [286], while long-range interactions are accounted for by the AC method. In this study, AC variables represented information on interactions between one amino acid and its 30 neighbouring amino acids in the protein sequence.	The model achieved a low prediction accuracy of 58.42% in a negative dataset created using the Prcp method [286].
UNISPPI [2013] [284]	This method used a decision tree model, predicting PPIs using only 20 amino acid frequency combinations from interacting and noninteracting proteins as learning features. This study indicated that asparagine, cysteine, and isoleucine frequencies are important features for discerning between interacting and noninteracting protein pairs.	This method was scalable due to using a limited number of attributes. Moreover, this method was based on experimentally validated instances from various species, covering many species.	Instances with a classification score of 0.50 were classified as neither PPIs nor non-PPIs, limiting the applicability of this method. Additionally, the obtained accuracy of 79.4% for interacting and 72.6% for noninteracting pairs are relatively low.
SVM-based method [2015] [285]	PPI prediction was addressed by integrating a support vector machine (SVM) and a novel matrix-based representation of the sequence order and dipeptide information of the primary protein sequence, extracting more information than amino acid dipeptide composition. The SVM classified the interaction between protein pairs using these feature vectors.	This method extracts more information hidden in protein primary sequences than amino acid dipeptide composition.	SVM algorithms performed relatively poorly with noisy data and are unsuitable for large datasets since training time may increase significantly [303]. Moreover, finding a proper kernel function was difficult.
DeepPPI [2017] [292]	The DeepPPI method used a deep neural network architecture network for each protein to extract high-level discriminative features from common protein descriptors. The interaction between two proteins was determined using the one-hot encoding label. This method comprises two different architectures: DeepPPI-Sep, which uses two separate networks as input for each protein, and DeepPPI-Con, which directly links two proteins in a single network.	This method can capture informative features of protein pairs by a layer-wise abstraction. In addition, DeepPPI can automatically learn an internal distributed feature representation from the data.	The accuracy of DeepPPI for All Human/ Yeast dataset are relatively low, and the accuracy of methods proposed in [304] exceeds that of DeepPPI.
Stacked Autoencoder (SAE) [2017] [18]	This method used a stacked autoencoder to predict PPI. The feature extraction from protein sequences was performed using autocovariance (AC) and the conjoint triad (CT).	SAE can learn hidden interaction features of protein sequences.	They used a synthetic negative interaction dataset, and the accuracy of this model for negative interactions is relatively low.
DPPI [2018] [298]	This method performed sequence-based PPI prediction using a deep, Siamese-like convolutional neural network combined with random projection and data augmentation. This method captured the composition information, sequential order of amino acids, and co-occurrence of interacting sequence motifs in a protein pair. Each protein was characterised as a probabilistic sequence profile generated by PASI-BLAST. The patterns in each sequence were identified using the convolutional module, comprising multiple layers. The representations learned by the convolutional module were projected to two different spaces using the random projection module, allowing DPPI to explore the combination of protein motifs.	DPPI addresses interactions for both homodimeric and heterodimeric proteins. Moreover, this method could model binding affinities.	This method yields lower PPI prediction accuracy on the S.cerevisiae core dataset from PIPR based on 5-fold cross-validation compared to PIPR [299] and DeepTrio [302].

Framework	Description	Advantage	Disadvantage
PIPR [2019] [299]	This study proposed an end-to-end framework for PPI prediction based on amino acid sequences using a deep residual recurrent convolutional neural network in the Siamese architecture. This method leveraged an automatic multi-granular feature selection to capture local significant and sequential features from protein sequences.	The Siamese-based learning architecture captured the mutual influence of protein pairs and allowed for generalising to address different PPI prediction tasks without needing predefined features.	RCNN was built using bidirectional gated recurrent units (bidirectional-GRU). However, GRUs suffer from slow convergence and low learning efficiency [305].
S-VGAE [2020] [300]	This model proposed a signed variational graph autoencoder (S-VGAE) that combined sequence information and graph structure. In this method, the PPI network was regarded as an undirected graph. This framework comprised three parts. First, coding the raw protein sequences. Second, the S-VGAE model extracted vector embedding for each protein with sequence information and graph structure. Finally, a simple three-layer softmax classifier. This model was inspired by the variational graph autoencoder (VGAE) [306] that uses latent variables to learn interpretable representations for undirected graphs.	In this method, the cost function was modified only to consider highly confident interactions, making it more robust to noise.	This model used the conjoint triad (CT) method [286] to encode amino acids. However, CT does not account for long-range interactions in the protein sequence.
ACT-SVM [2020] [289]	This method performed feature extraction on the protein sequence to obtain a vector, composition, and transition descriptor and integrated them into a vector. Then, the feature vector was fed into the SVM classifier. The performance of this method was evaluated using 5-fold, 8-fold, and 10-fold cross-validation on H. pylori and human datasets.	They have observed that SVM method outperforms K-Nearest Neighbour (KNN), ANN, RFM, Naive Bayes, Logistic Regression, s for the H. pylori protein pairs.	Finding the proper kernel and hyperparameters was challenging, and training time for SVM classifiers increases with dataset size [307].
D-SCRIPT [2021] [165]	Deep sequence contact residue interaction prediction transfer (D-script) is an interpretable deep learning method generating structurally informative features given protein sequences using a pre-trained language model from [290]. This method used projection modules to reduce the dimension of features, including the residue-contact map of the protein. Finally, the interaction probability was predicted based on the contact maps.	D-SCRIPT generalised to new species considering the sparsity of training data for most model organisms (i.e., it was relatively accurate for cross-species PPI prediction).	Despite its performance for cross-species PPI prediction, D-SCRIPT underperformed on within-species evaluations. The training dataset only included proteins with 50–800 amino acids, limiting the applicability of this method.
SPNet [2021] [9]	The Siamese pyramid network (SPNet) architecture used self-binding and folding amino acid sequences to predict the binding probability for two proteins based solely on their amino acid sequences. Subsequent screening through potential candidates was performed based on binding probabilities.	This architecture consisted of a multilevel pyramid feature structure encompassing various PPI mechanisms to reduce gradient explosion and disappearance, a multilevel Siamese neural network with an attention mechanism, and a multilevel, trainable binding probability prediction network.	
BiLSTM-RF [2021] [291]	The BiLSTM-RF model extracted features of protein pairs in the human database. BiLSTM comprises forward and backwards LSTMs and is capable of bidirectional encoding (i.e., encoding front-to-back and back-to-front information). A random forest classifier (RF) was built with 100 trees and used a voting strategy to integrate these results to predict the interaction.	BiLSTM extracted the sequence and position of the biological information in the protein sequence.	LSTM models are computationally demanding and slow. Moreover, a large number of trees in the random forest leads to a longer training time.
Heterogeneous Network [2021] [296]	PPI prediction was performed using a computational sequence and network representation learning-based model. Local features were extracted from the protein sequence using the *k*-mer method (k = 3), while global features were extracted from the heterogeneous network. The latter captured network structure and obtained potential linked information. This method integrated local features with global features to represent protein nodes.	The protein node contained protein attribute and network structure information by integrating local and global features.	Model accuracy is relatively low compared to other deep learning methods such as DPPI [298].

Framework	Description	Advantage	Disadvantage
OR-RCNN [2021] [301]	This method was called ordinal regression and recurrent convolutional neural network (OR-RCNN), which predicted PPIs based on their confidence score. The architecture comprised two recurrent convolutional neural networks (RCNNs) encoders, which shared the same parameters, to extract robust local features and sequential information from protein pairs. Then, one novel embedding vector was obtained by element-wise multiplication of the two embedding vectors from RCNNs. The second part of the architecture performed an ordinal regression model via multiple sub-classifiers that use the ordinal information behind the confidence score. Finally, the confidence score determined the existence of PPI with a threshold.	This method offered better accuracy compared to some existing models, such as autocovariance [288] and composition transition distribution (CTD) descriptor [308] for feature description, and random forest (RF) [309], extreme gradient boosting (XGBoost) [310], and support vector machine (SVM) [311] for the prediction.	The RCNN was built using bidirectional gated recurrent units (bidirectional-GRU). However, GRUs suffer from slow convergence and low learning efficiency [305].
DeepTrio [2022] [302]	The DeepTrio method used a deep-learning framework based on a mask multiscale CNN architecture that performed binary PPI prediction by capturing multiscale contextual information of protein sequences using multiple parallel filters. This method used a single-protein class, allowing it to distinguish relative and intrinsic properties. This method was also made available as an online tool to address cross-platform usage and dependency-related issues.	DeepTrio is available both online and offline.	DeepTrio yields lower PPI prediction accuracy on the S.cerevisiae core dataset from PIPR based on 5-fold cross-validation compared to PIPR [299]. DeepTrio achieves lower PPI prediction accuracy on the S.cerevisiae core dataset from DeepFE-PPI based on 5-fold cross-validation compared to DeepFE-PPI [312].

Journal of Pharmaceutical Sciences 110 (2021) 665-681

Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Review

Prediction Machines: Applied Machine Learning for Therapeutic Protein Design and Development

Tim J. Kamerzell [a, b, *], C. Russell Middaugh [a]

[a] Department of Pharmaceutical Chemistry, The University of Kansas, Lawrence, KS, USA
[b] Division of Internal Medicine, HCA MidWest Health, Overland Park, KS, USA

* Source: xxxx, 2022.0x.xx

Characteristics of Machine Learning Models Used in Pharmaceutical Protein Development.

	Machine Learning Algorithms and Tasks	Types of Data Input and Features	Perf	Ref
Asparagine Aspartic Acid Degradation	RF Classification and Regression	• Dataset of n = 776 mAb peptides. Training set includes 64 IgG1s and 3 IgG4s. Validation set contains 10 IgG1s and 2 IgG4s • Binary classifier for deamidation lability and regression analysis to predict deamidation rate • Features include 12 descriptors for primary sequence (2), backbone orientation (3), side chain orientation (2), solvent accessibility (2) and hydrogen bonding (3)	Acc 100% R^2 0.93	60
	SVM, RF,NB,KNN,ANN,PLS Classification	• Training set consists of 194 Asn residues and 25 proteins. Test set consists of 81 Asn residues and 3 protein structures • Binary classifier for deamidation lability • Descriptors include one experimental measure (deamidation half-life) and several structural features including 3D structure based properties, backbone orientation, side chain orientation, local secondary structure, B-factors, solvent accessibility and reaction coordinate • Feature selection with recursive feature elimination • 10 fold cross validation • Data sets unbalanced toward the negative case. Therefore used recall and specificity for further characterization	Acc 95% AUC 0.96 Recall 0.80 Spec 0.96	59
	SVM,DLANN,DT, RP Classification	• Dataset of 37 mAbs including 55 Asn hotspots and 940 Asn non hotspots, 40 Asp hotspots and 1425 Asp non hotspots. • Binary classifier for hotspot • Features include 20 descriptors, secondary structure, transition state accessibility, nucleophilicity, pKa, solvent accessibility • Dataset unbalanced toward non hotspots therefore a standard weighting scheme using inverse of class frequency employed • Monte Carlo cross validation used	TPR 0.94	62
	DT Classification	• Dataset of 5 IgG1 and 5 IgG4 mAbs • Classification 4 groups with increasing deamidation propensity • Features include backbone flexibility, C–N distance and solvent accessibility • No clear validation method or performance metrics provided		63
Protein Oxidation	SVM,RF,NN Classification	• Dataset of 113 polypeptides containing 975 Met residues of which 122 oxidation prone and 853 oxidation resistant • Binary classifier for oxidation • 54 Features include solvent accessibility, primary structure, met-aromatic residue distance, 3D structure • Feature extraction and selection with mRMR and GI • 10 fold cross validation • Stratified random sampling used due to unbalanced data	Acc 84% Spec 0.87 F-meas 0.48	64
	SVM, RF, NN, FDA Classification	• 1646 proteins with 2616 methionine sulfoxides • Binary for oxidized • 54 Features including 40 primary structure, 14 tertiary structure, 2 solvent accessibility, 2 entropy and 1 frequency feature • Feature extraction and selection with AI and GI • 10 fold cross validation • Stratified random sampling used due to unbalanced data	Acc 82% AUC 0.85 Sens 77 Spec 82	105
	RF, MLR Classification and Regression	• 172 Met residues across 122 mAbs • Binary for oxidized and RF regression model trained to predict relative change in % oxidized species • 18 descriptors initially evaluated. Final model included 2 structural and 2 dynamic features including the number of overlaps between atoms of Met residue with neighbors, solvent accessibility, and mean square fluctuation of Met C alpha atom • Feature extraction and selection with GI • Hold out validation	R 0.94 Acc 96% AUC 0.96 Spec 0.98 Sens 0.83	67
	NN Classification	• 166 peptide fragments of which 32 are positive and 134 negativ • Binary for oxidized • Initially 735 features which was decreased to the top 16 features including secondary structure, solvent accessibility, disorder, amino acid properties and side chain count of carbon atom deviation from mean • Feature selection with mRMR • Jackknife cross validation	Acc 92% Sens 0.84 Spec 0.94	66
Protein Solution Viscosity	Log linear, PCR Regression	• 14 mAbs at a single high concentration • Features include Fv net charge, Fv charge symmetry and Fv hydrophobicity • Leave one out cross validation	R 0.8	55
	Log linear, stepwise linear Regression	• viscosity from concentration dependence of 16 mAbs • Both log linear and stepwise linear regression used • Features include sequence structural attributes including charge, pI, zeta potential, dipole moment, solvent accessibility and nonpolar surface • Slope of log linear equation (B) and stepwise linear approach used to identify correlations and predict B • leave one out cross validatoin	varies	54

T.J. Kamerzell, C.R. Middaugh / Journal of Pharmaceutical Sciences 110 (2021) 665–681

* Source: xxxx, 2022.0x.xx

	Machine Learning Algorithms and Tasks	Types of Data Input and Features	Perf	Ref
	Log linear Regression	• Viscosity from concentration dependence of 11 mAbs • Slope of log linear equation and stepwise linear regression used • Features include net charge, zeta potential, isoelectric point, hydrophilic and hydrophobic surface area and Kyte-Doolittle hydrophobicity moment • leave one out cross validation	R^2 0.88 -0.96	53
	ANN, SVM, ensemble, DT, NB, KNN,LDA, PCA Classification and Regression	The authors in this review used the viscosity dependent data and features provided in references[51,53] for both classification and regression of protein solution viscosity. For a detailed discussion of the models and parameters used see manuscript text and figure legends	varies	TJK CRM
Sub-visible Particles	KLD, MDS, clustering Classification	• Classification of particle type from flow microscopy images after freeze thaw stress, shaking, pH stress, thermal stress or formulation • 3 samples per stress and two replicates per sample measured by flow microscopy • number of particles detected by flow microscopy in all the samples ranged from 198 to 45,431 with a median of 21,062 • Divergence calculated using particle descriptors including size and aspect ratio • Matrix of divergences projected using MDS	Acc 100%	92,94
	CNN, SVM, KNN, DT, ensemble Classification	• 100,000 particles per stress source which includes friability, freeze-thawing, agitation and heating • Feature extraction using PCA on 37 texture metrics from particles used to train classifiers • images also used as direct inputs in training CNN classifiers • n = 600 randomly selected individual particles with a sampling frequency, f = 20 • 10 fold cross validation	Acc 95–99%	93
	CNN Classification	• 2×10^5 total training samples • CNN to identify aggregation inducing stress which includes freeze thaw stress, agitation stress, pump recirculation and protein and silicone oil mixtures • CNN was three layers and a total of 28,640 trainable parameters	Acc 100%	95
	RF, logistic regression Classification	• Protein particle concentrations ~50,000/mL • Random Forest used to classify and count silicone oil and non-silicone oil particles in formulations • Features include the aspect ratio, circularity, intensity mean, intensity standard deviation, intensity minimum and intensity maximum • Random forest filters compared to Logistic S-factor filters • k-fold cross validation	varies	96
Biophysical Stability	RF Classification	• Random Forest used to classify acid stable and neutral proteins • 393 proteins and 889 sequence based features • GI used to rank feature importance. Top 10–25 features used • five-fold cross validation	Acc 75% AUC 0.91	86
	Bagging, LDA, PCA, PLS Classification and Regression	• Raman spectroscopy used to classify four mAbs and predict concentration • Crowdsourced machine learning analyses • Collected ~350 spectra for each of four mAbs • 10 fold cross-validation	varies	106
	kNN, SVM, LDA, QDA, NB, DT, RF, AdaBoost, MI Classification	• Bio-similarity assessment of 4 mAbs in 2 different formulations resulting in 8 samples in triplicate for total of 24 samples • 2066 features from 4 analytical methods • MI for feature ranking and PCA for visualization • Parameter optimization with grid search • Leave one and leave 8 out cross-validation	Acc 100%	76
	ANN, PLS Regression and Classification	• 6 mAbs and 24 conditions per protein resulting in 144 samples • Only number of each amino acid species of the protein used as input parameter • light scattering and fluorescence used to measure melting temperature (Tm), aggregation (Tagg) and kD • ANN and PLS for prediction of Tm, Tagg	R^2 0.98 0.94	84
	kNN, SVM, DT, AdaBoost, RF, NB, LDA, PCA, MI Classification	• 35 samples • Multiple biophysical techniques including UV–Vis absorption spectroscopy, FTIR, CD and HPLC used to characterize Crofelemer • Approximately 7000 features per sample • PCA for feature scaling and MI to identify important features • Monte Carlo cross-validation	Acc 99%	82

* Source: xxxx, 2022.0x.xx

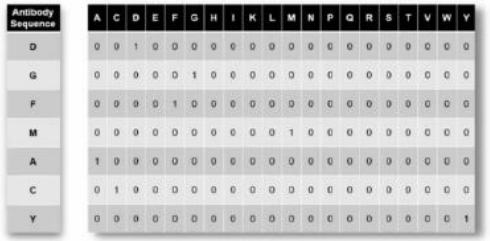
OXFORD

Briefings in Bioinformatics, 2022, **23**(**4**), 1–20
https://doi.org/10.1093/bib/bbac267
Advance access publication date: 14 July 2022
Review

Machine-designed biotherapeutics: opportunities, feasibility and advantages of deep learning in computational antibody discovery

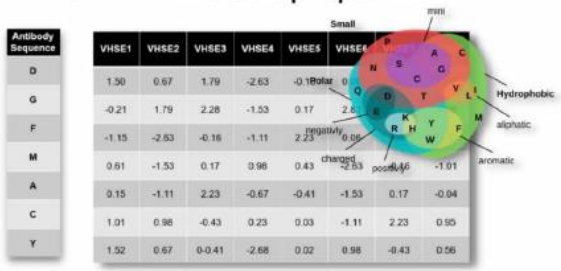
Wiktoria Wilman, Sonia Wróbel, Weronika Bielska, Piotr Deszynski, Paweł Dudzic, Igor Jaszczyszyn, Jędrzej Kaniewski, Jakub Młokosiewicz, Anahita Rouyan, Tadeusz Satława, Sandeep Kumar, Victor Greiff and Konrad Krawczyk
Corresponding author. Dr Konrad Krawczyk, NaturalAntibody. E-mail: konrad@naturalantibody.com, konrad@naturalantibody.com

A. One-hot encoding

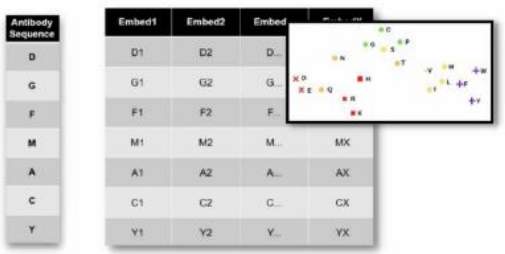


B. Substitution matrix-based

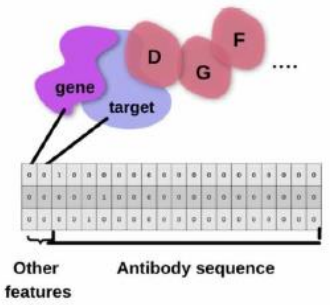
C. Amino acids properties



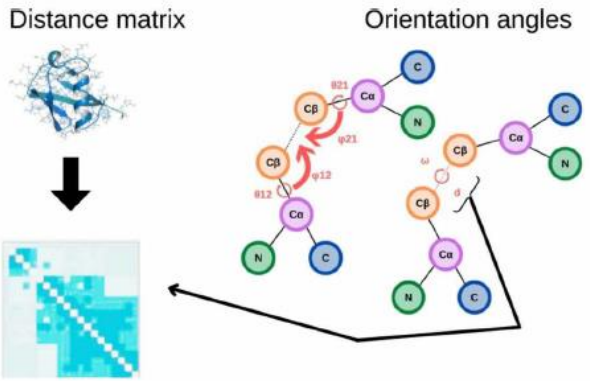
D. Learned amino acids properties



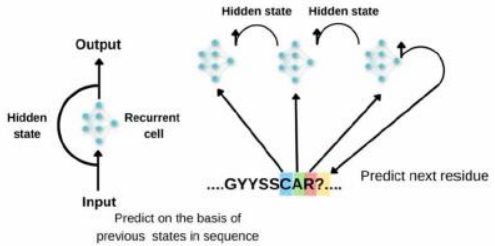
E. Encoding of suplementary attributes



F. Encoding of structural features

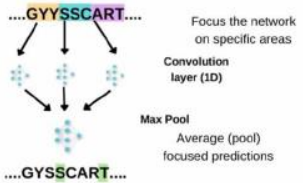


A. Recurrent Neural Networks



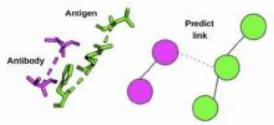
Predict on the basis of previous states in sequence

B. Convolutional Neural Networks



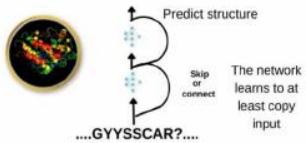
Predict paratope

C. Graph Neural Networks

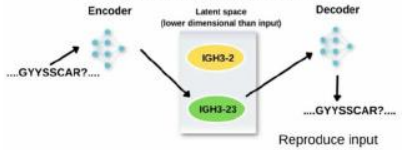


Introduce meaningful connections between input nodes and discover new links.

D. Residual Neural Network



E. Encoder-decoder architecture

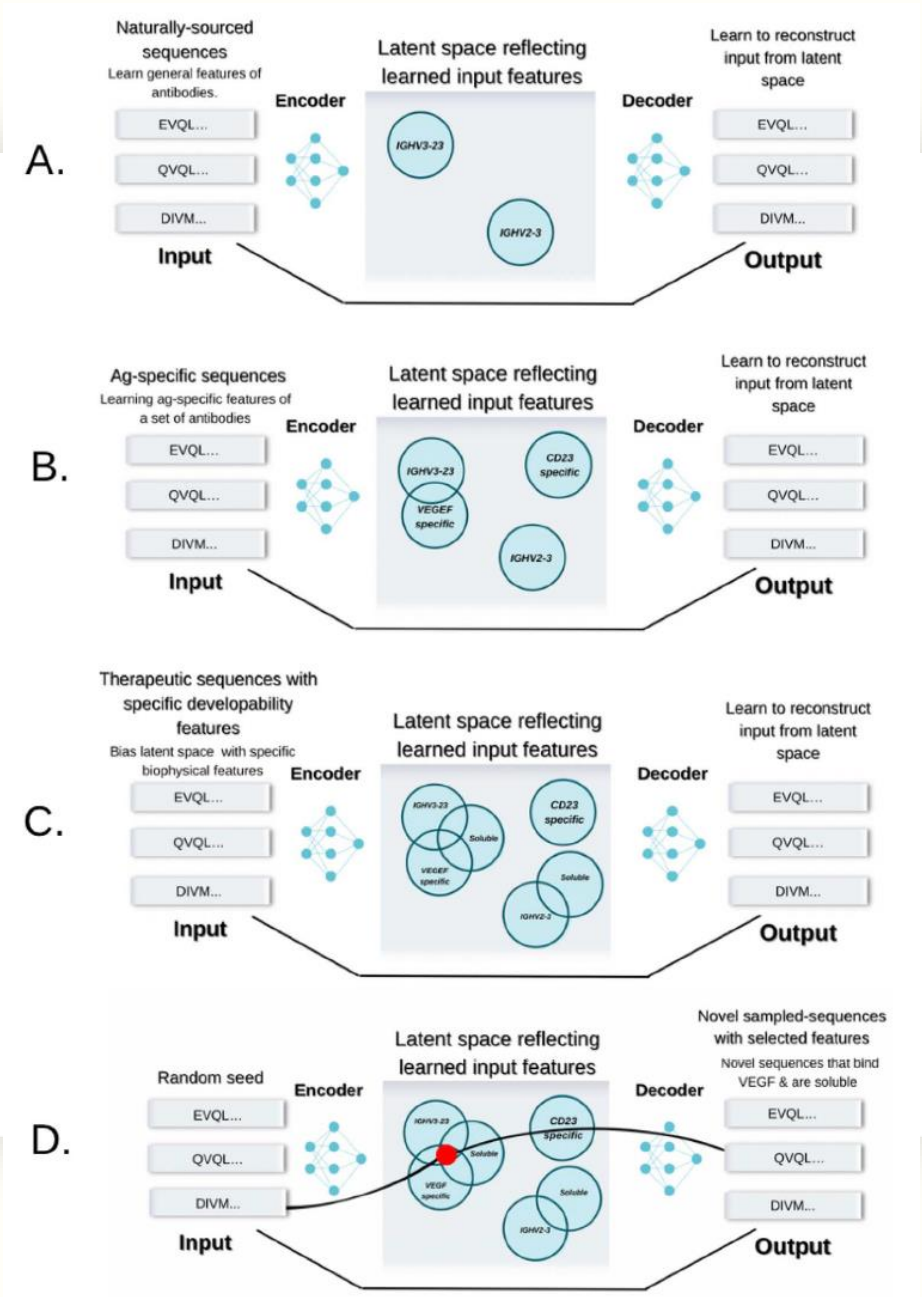


Reproduce input

Category	Method	Problem solved	Training input	Architecture	Training parameters	Libraries	Availability	Paper
Structure prediction	DeepH3	CDRH3 prediction	1388 structures	Series of 1D and 2D convolutions (3 1D + 25 2D blocks)	30 epochs, batch size 4, 35 h using one NVIDIA Tesla K80 Graphics processing unit (GPU)	PyTorch	link	[27]
	DeepAb	V region structure prediction	118 386 sequences and 1692 structures	A 1D ResNet (1D convolution followed by three 1D ResNet blocks) and the bi-LSTM encoder	60 epochs, batch size 128, NVIDIA K80 GPU requiring 60 h	PyTorch	link	[19]
	AbLooper	CDR Prediction	3438 structures	Five E(n)-equivariant graph neural networks (EGNNs), each one with four layers	NVIDIA Tesla V100 GPU, predict the CDRs for one hundred structures in under five seconds	PyTorch	link	[40]
	NanoNet	Heavy chain prediction	~2000 structures	Two 1D ResNets with input tensor of 140 × 22	batch size of 16 and ~ 130 epochs,10 min on a GeForce RTX 2080 Ti	Keras/TensorFlow	link	[18]*
Humanization/Deimmunization	Nativeness LSTM	Learn distribution of amino acids at positions	400 000 sequences	Bidirectional LSTM with dimensionality 64	10 epochs	PyTorch	link	[31]
	Sapiens	Antibody humanization	20 milion heavy chains and 19 milion light chains	RoBERTA transformer, 4 layers, 8 attention heads, 568 857 parameters	700 epochs for heavy chains, 300 epochs for light chains	PyTorch/-Fairseq [41]	link	[24]
	hu-Mab	Discriminate between human/mouse sequences	65 million sequences with 13 million non-human ones	Random Forest	n/a	scikit-learn	link	[42]
Binding models	Parapred	Paratope residues prediction	1662 sequences (277 antibody–antigen complexes × 6 Complementarity determining regions each) and tested on the same dataset using 10-fold cross-validation technique	Convolutional and recurrent neural networks	16 epochs, 32 batch size	Keras	link	[43]
	Epitope3d	Conformational epitopes prediction	1351 antibody–antigen structures (covering 40 842 epitope residues) and 180 unbound antigen structures; tested on 20 unbound antigen structures; 45 unbound antigen structures used for external blind test	Supervised learning algorithms: Multi-layer Perceptron, Support Vector Machines, K-Nearest Neighbor, Adaboost, Gaussian processes (GP), Random Forest, Gradient Boost, XGBoost, Extra Trees	N/A	scikit-learn Python	link	[44]
	mmCSM-AB	Prediction of the consequences of multiple point mutations on antibody–antigen binding affinity	1640 mutations with associated changes in binding affinity (905 single missense mutations and 735 modeled reverse mutations); tested on 242 multiple missense mutations with associated changes in binding affinity	Supervised learning algorithms for example: Random Forest, Extra Trees, Gradient Boost, XGBoost, SVM and Gaussian Process	n/a	scikit-learn Python	link	[45]

Category	Method	Problem solved	Training input	Architecture	Training parameters	Libraries	Availability	Paper
	Phage display LSTM	Generate novel kynurenine binding sequences from LSTM	959 sequences	LSTM, two layers with 64 units.	269 epochs	Keras/Tensorflow	n/a	[46]
	Phage display CNN	Predict phage enrichment and generate novel CDRH3	96 847 sequences (largest dataset on github)	Ensemble of CNNs, largest with two convolutional layers and 18 706 parameters	20 epochs	Keras	link	[47]
	Image-based prediction	Distinguish between binding antibodies and lineages	24 953 models with calculated fingerprints from 308 EBOV and 54 HIV antibodies.	ResNet-50 [48]	Pre-trained model	Keras/Tensorflow	link	[33]
	Paratope and Epitope Prediction with graph Convolution Attention Network (PECAN)	Epitope and paratope prediction	162 structures for epitope prediction and 460 for paratope prediction	Graph Convolutional Attention Network	Up to 250 epochs, batch size of 32 (multiple parameters tested)	Tensorflow	link	[20]
	DLAB	Sorting of protein docking poses	759 Antibody–antigen complexes	Convolutional Neural Network	n/a	PyTorch	link	[32]
Embeddings/Language Methods	immune2vec	Embed CDRH3 into 100 dimensions using skip-gram	15,63 million sequences	Two dense layers	n/a	Gensim	link	[49]
	ProtVec CDRH3	Embed CDRH3 sequences to predict COVID-19 status	COVID-119 data from OAS	Based on ProtVec from Harvard DataVerse [50] and SVM	Reused previous model.	Reused previous model	link	[51]*
	AntiBerty	Masked language modeling, paratope prediction	558 milion sequences	BERT transformer encoder model, 8 layers, 26 M trainable parameters.	8 epochs, 10 days on four NVIDIA A100 GPUs	PyTorch	n/a	[39]*
	AntiBerta	Masked language modeling, paratope prediction	57 million sequences	Antibody-specific Bi-directional Encoder Representation from Transformers, 86 m parameters	12-layer transformer model that is pre-trained on 57 M human BCR sequences, 3 epochs, batch size of 96 across 8 NVIDIA V100 GPUs	PyTorch	n/a	[25]
	AbLang	Masked language modeling, reconstruct erroneous sequences	14 milion heavy chains, 200 000 light chains training. Evaluation sets of 100 k, 50 k for heavy lights respectively.	Based on RoBERTA from HuggingFace. 12 layers.	20 epochs for heavy chains, batch 8192, light chains 40 epochs 4096 batch size	PyTorch	link	[52]*
Generative methods/antibody design	Mouse VAE	Model latent space of CDR triples of antigen challenged mice	243 374 sequences.	VAE with encoder and decoder each having two dense layers (256 512 units each)	200 Epochs on a single GPU from the ETH cluster.	Tensorflow	n/a (available after peer review)	[53]*
	Developability-controlled GAN	Learn latent representation of human sequences and bias it towards biophysical properties	400 000 sequences	Generative Adversarial Network, (single chain) seven layers consisting of 2D convolution and dense layers.	500 epochs, batch size of 128	Keras/Tensorflow	n/a	[35]*
	Nanobody generation	Autoregression on nanobody sequences to generate novel CDRH3	1.2 milion sequences	ResNet with nine blocks with six dilated convolutional layers.	250 000 updates, batch size of 30.	Tensorflow/PyTorch	link	[17]

Category	Method	Problem solved	Training input	Architecture	Training parameters	Libraries	Availability	Paper
	In silico LSTM	*In silico* proof-of-principle of virtually unconstrained antigen-specific antibody sequence generation	70 000 murine CDR3 sequences	1024 LSTM with embedding layer and dense output layer.	20 epochs, batch size 64	Tensorflow	link	[54]
	Immunoglobulin Language Model (IgLM)	Masked language modeling, generate synthetic libraries of antibodies by solving masked language model	558 milion sequences	Transformer decoder architecture based on the GPT-2 model with 512 embeddings, 12 milion parameters	batch size of 512 and 2 gradient accumulation steps using DeepSpeed, 3 days when distributed across 4 NVIDIA A100 GPUs	GPT-2 from Hugging-Face	n/a	[55]*
	IG-VAE	Immunoglobulins structure generation	10 768 immunoglobulins structures (including 4154 non-sequence-redundant structures)- set covers almost 100% of the antibody structure database (AbDb); Tested on 5000 structures from the latent space of the Ig-VAE	VAE	n/a	PyTorch	n/a	[34]*
	Generative method Benchmarking: (AR) the sequence-based autoregressive generative model, geometric vector perceptron (GVP) the precise structure-based graph neural network and (Fold2Seq) fold-based generative model	Antibody CDR regions design based on portion of sequence or structure.	Sequences from natural llama nanobody repertoire	AR- Autoregressive Causal Dilated Convolutions; GVP -based Encoder-Decoder GNN; Fold2Seq-Encoder-Decoder Transformer	n/a	n/a	n/a	[56]*
	GNN-based generation	CDRs sequence and 3D structure design	~5000 structures. For CDR-H1, the train/validation/test size is 4050, 359 and 326. For CDR-H2, the train/validation/test size is 3876, 483 and 376. For CDRH3, the train/validation/test size is 3896, 403 and 437.	Message passing network (MPN): Iterative Refinement Graph Neural Network (RefineGNN)	batch size of 16, dropout of 0.2 and learning rate of 0.0005	n/a	n/a	[23]*
	AntBO	CDRH3 region design		Bayesian Optimization and GP	87 cores 12 GB GPU memory	GPyTorch, Botorch	n/a	[57]*


A.
Naturally-sourced sequences
Learn general features of antibodies.
EVQL...
QVQL...
DIVM...
Input
Encoder
Latent space reflecting learned input features
IGHV3-23
IGHV2-3
Decoder
Learn to reconstruct input from latent space
EVQL...
QVQL...
DIVM...
Output
B.
Ag-specific sequences
Learning ag-specific features of a set of antibodies
EVQL...
QVQL...
DIVM...
Input
Encoder
Latent space reflecting learned input features
IGHV3-23
CD23 specific
VEGEF specific
IGHV2-3
Decoder
Learn to reconstruct input from latent space
EVQL...
QVQL...
DIVM...
Output
C.
Therapeutic sequences with specific developability features
Bias latent space with specific biophysical features
EVQL...
QVQL...
DIVM...
Input
Encoder
Latent space reflecting learned input features
IGHV3-23
Soluble
CD23 specific
VEGEF specific
Soluble
IGHV2-3
Decoder
Learn to reconstruct input from latent space
EVQL...
QVQL...
DIVM...
Output
D.
Random seed
EVQL...
QVQL...
DIVM...
Input
Encoder
Latent space reflecting learned input features
IGHV3-23
Soluble
CD23 specific
VEGF specific
Soluble
IGHV2-3
Decoder
Novel sampled-sequences with selected features
Novel sequences that bind VEGF & are soluble
EVQL...
QVQL...
DIVM...
Output

Antiverse is a new type of antibody discovery company accelerating drug development. The Antiverse platform exists at the intersection of structural biology, machine learning and medicine to enable breakthroughs to happen more quickly and cost-effectively.

Antiverse **prevents diversity loss** during amplification to uncover more diverse and rare antibodies.

Traditional in vitro screening:

10^{10} antibodies

↓

3 amplification rounds

↓

10 antibodies

Antiverse discovery:

Antigen-antibody database

↓

AI-augmented screening

↓

96 antibodies

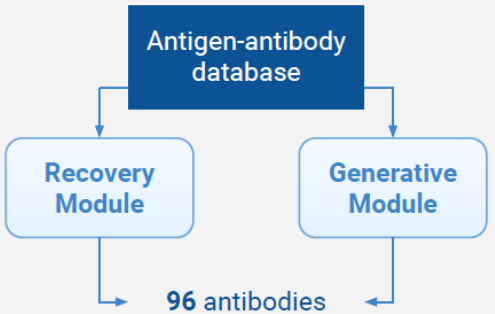
Library Diversity

Antiverse

Exploring the Full Functional Spacing

Antiverse **provides more candidates** by analysing NGS data, clustering on multi-dimensional space, and selecting based on sequential and structural grouping. The generative module **offers new sequences** and gives you options that haven't even been considered.

Antiverse AI-Augmented Discovery:



Antiverse

Antiverse is recognized as one of the top biotech startups in the UK with our antibody discovery service already in use by big pharma. The main feature of the company is **10x Diversity with AI-Augmented Drug Discovery**.

Existing antibody discovery methods are well-developed and often effective at discovering binders. But when there is a need to find the best possible candidate, or when finding a suitable candidate is hard with current methods, the options are **limited** and often **costly**.

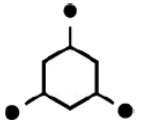
Antiverse uses **next-generation sequencing (NGS)** to extract more data from existing workloads. The **AI-Augmented Drug Discovery platform** and trained models analyse the statistics gained from thousands of experiments. These outputs are compared against known data in order to select best candidates.

Target Selection

Antiverse provides targeted options in order to focus on testing safely once there are too many antibody-antigen binding options.

Binder Recovery

Antiverse can help to find sufficient potential binders that can be missed by conventional methods.

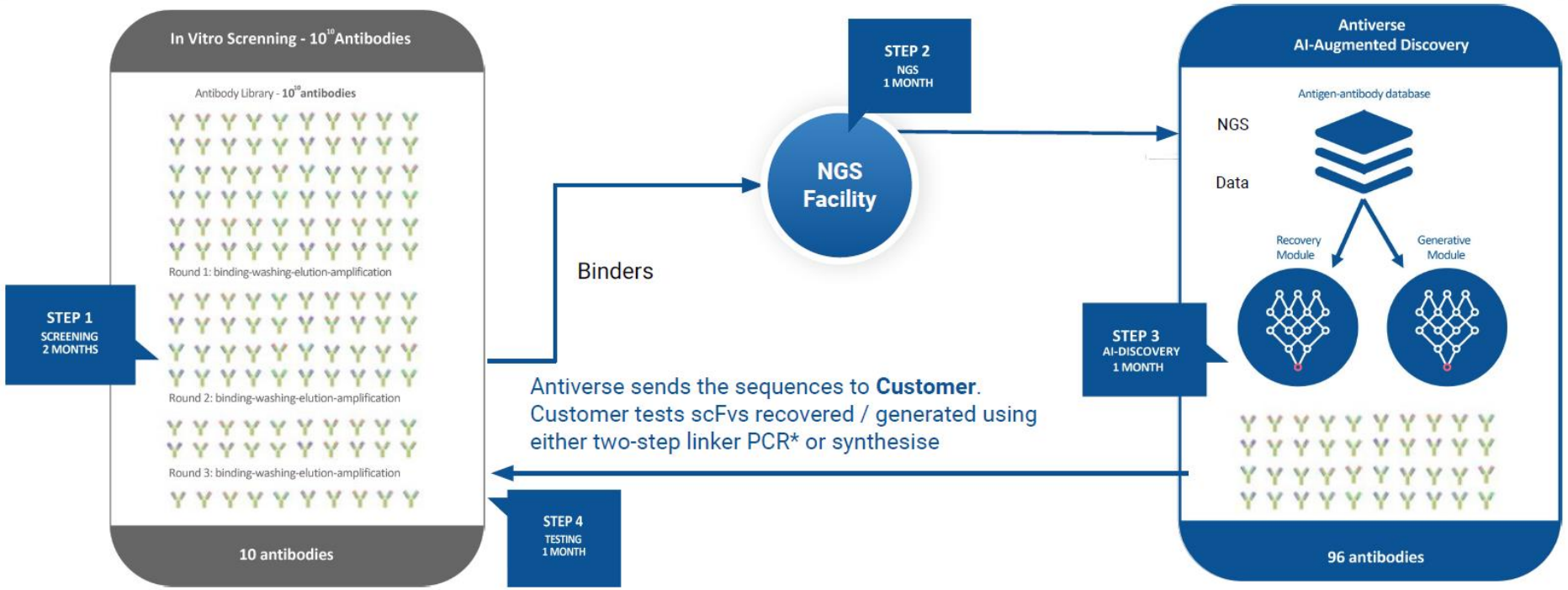


Binder Customisation

Antiverse can generate new binder variants that will be sufficient for clients purposes.

Antiverse

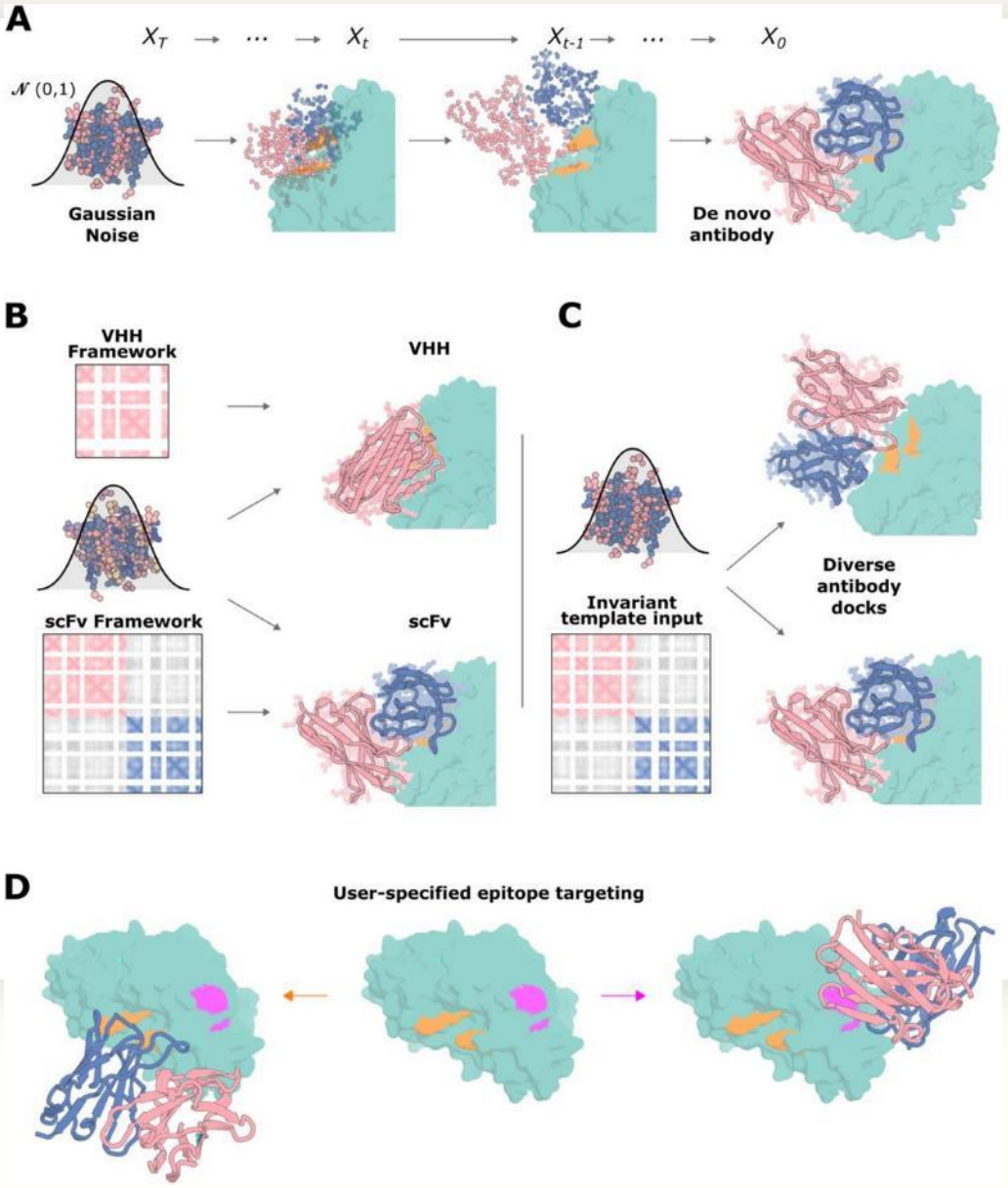
The **Antiverse AI-ADD** system found each and every cluster identified by other methods, plus more. These additional clusters contained rare and unique sequences.



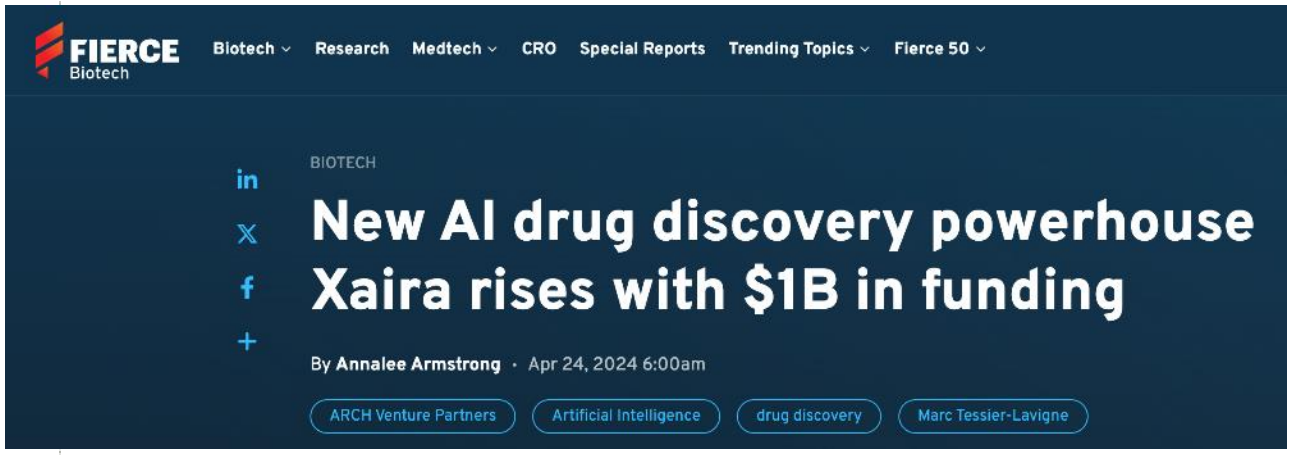
bioRxiv preprint doi: https://doi.org/10.1101/2024.03.14.585103; this version posted March 18, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Atomically accurate de novo design of single-domain antibodies

Nathaniel R. Bennett[‡1,2,3], Joseph L. Watson[*‡1,2], Robert J. Ragotte[‡1,2], Andrew J. Borst[‡1,2], Déjenaé L. See[#1,2,4], Connor Weidle[#1,2], Riti Biswas[1,2,3], Ellen L. Shrock[1,2], Philip J. Y. Leung[1,2,3], Buwei Huang[1,2,4], Inna Goreshnik[1,2,5], Russell Ault[6,7], Kenneth D. Carr[2], Benedikt Singer[1,2], Cameron Criswell[1,2], Dionne Vafeados[2], Mariana Garcia Sanchez[2], Ho Min Kim[8,9], Susana Vázquez Torres[1,2,10], Sidney Chan[2], David Baker[*1,2,5]


A
X_T
X_t
X_{t-1}
X_0
$\mathcal{N}$ (0,1)
Gaussian Noise
De novo antibody
B
VHH Framework
VHH
scFv Framework
scFv
C
Invariant template input
Diverse antibody docks
D
User-specified epitope targeting


nature
Explore content
About the journal
Publish with us
Subscribe
nature › news › article
NEWS | 19 March 2024
‘A landmark moment’: scientists use AI to design antibodies from scratch
Modified protein-design tool could make it easier to tackle challenging drug targets – but AI antibodies are still a long way from reaching the clinic.
By Ewen Callaway
Researchers have used generative artificial intelligence (AI) to help them make completely new antibodies for the first time.


FIERCE Biotech
Biotech
Research
Medtech
CRO
Special Reports
Trending Topics
Fierce 50
BIOTECH
New AI drug discovery powerhouse Xaira rises with $1B in funding
By Annalee Armstrong · Apr 24, 2024 6:00am
ARCH Venture Partners
Artificial Intelligence
drug discovery
Marc Tessier-Lavigne

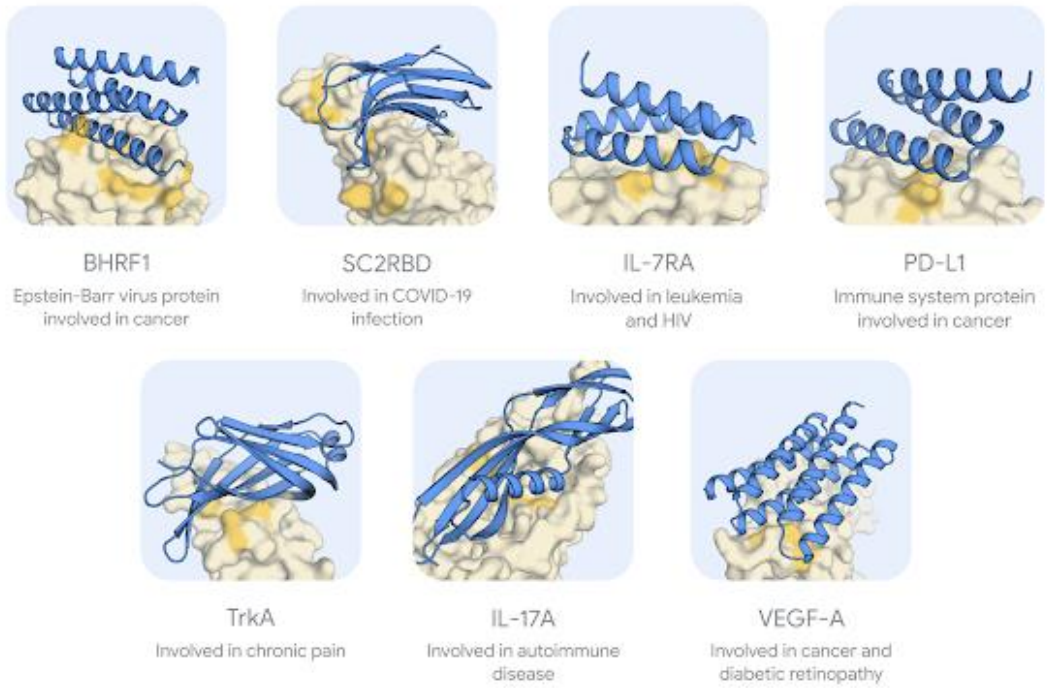
Google DeepMind

2024-09-05

De novo design of high-affinity protein binders with AlphaProteo

Vinicius Zambaldi[*,1], David La[*,1], Alexander E. Chu[*,1], Harshnira Patani[*,1], Amy E. Danson[*,1], Tristan O. C. Kwan[*,1], Thomas Frerix[*,1], Rosalia G. Schneider[*,1], David Saxton[*,1], Ashok Thillaisundaram[*,1], Zachary Wu[*,1], Isabel Moraes[2], Oskar Lange[2], Eliseo Papa[1], Gabriella Stanton[1], Victor Martin[1], Sukhdeep Singh[1], Lai H. Wong[1], Russ Bates[2], Simon A. Kohl[2], Josh Abramson[1], Andrew W. Senior[1], Yilmaz Alguel[3], Mary Y. Wu[4], Irene M. Aspalter[5], Katie Bentley[5,6], David L.V. Bauer[7], Peter Cherepanov[3], Demis Hassabis[1], Pushmeet Kohli[1], Rob Fergus[1,†] and Jue Wang[1,†]

[*]Equal contributions, [†]Equal supervision, [1]Google DeepMind, [2]Work performed while at Google DeepMind, [3]The Chromatin Structure and Mobile DNA Laboratory, The Francis Crick Institute, London, UK, [4]COVID Surveillance Unit, The Francis Crick Institute, London, UK, [5]Cellular Adaptive Behaviour Laboratory, The Francis Crick Institute, London, UK., [6]Department of Informatics, King's College London, London, UK. K.B. performed the work at the Cellular Adaptive Behaviour Laboratory, The Francis Crick Institute, London, UK, [7]RNA Virus Replication Laboratory, The Francis Crick Institute, London, UK



JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JACS

Perspective

Advancing Targeted Protein Degradation via Multiomics Profiling and Artificial Intelligence

Miquel Duran-Frigola,* Marko Cigler, and Georg E. Winter*

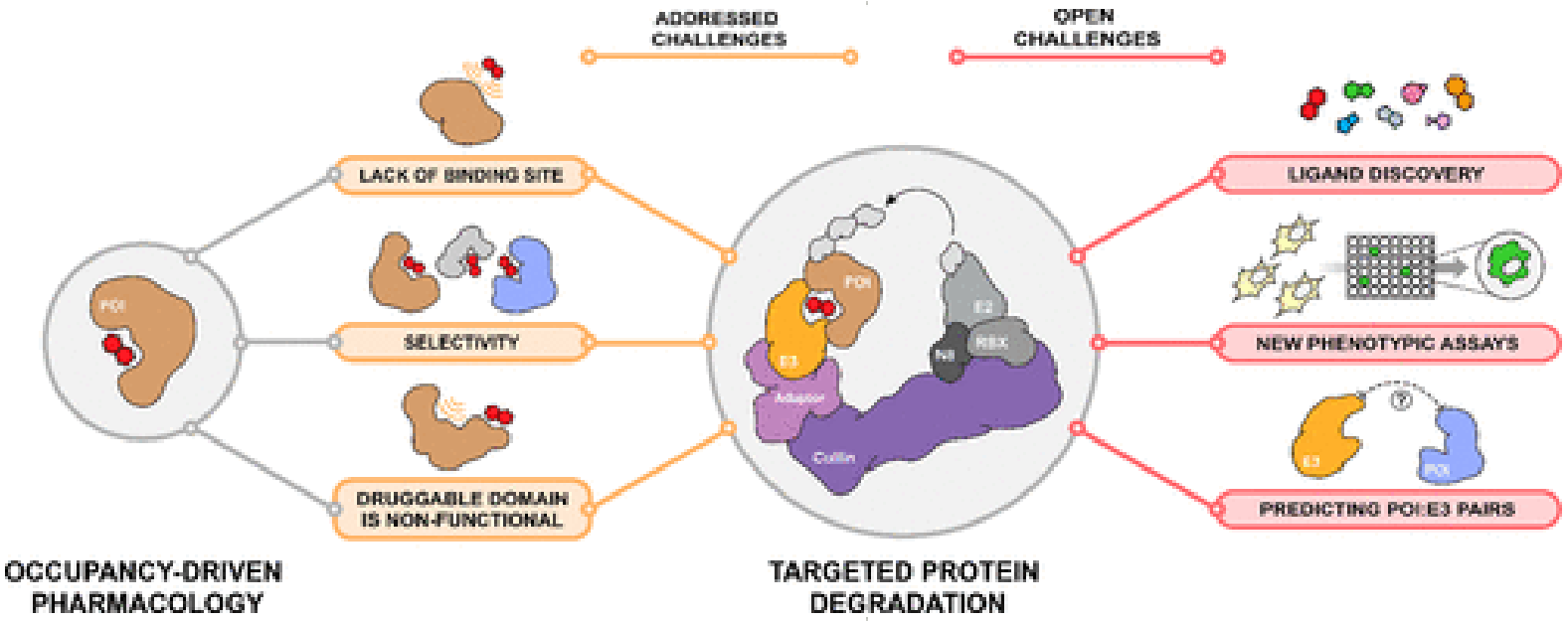
Cite This: *J. Am. Chem. Soc.* 2023, 145, 2711−2732

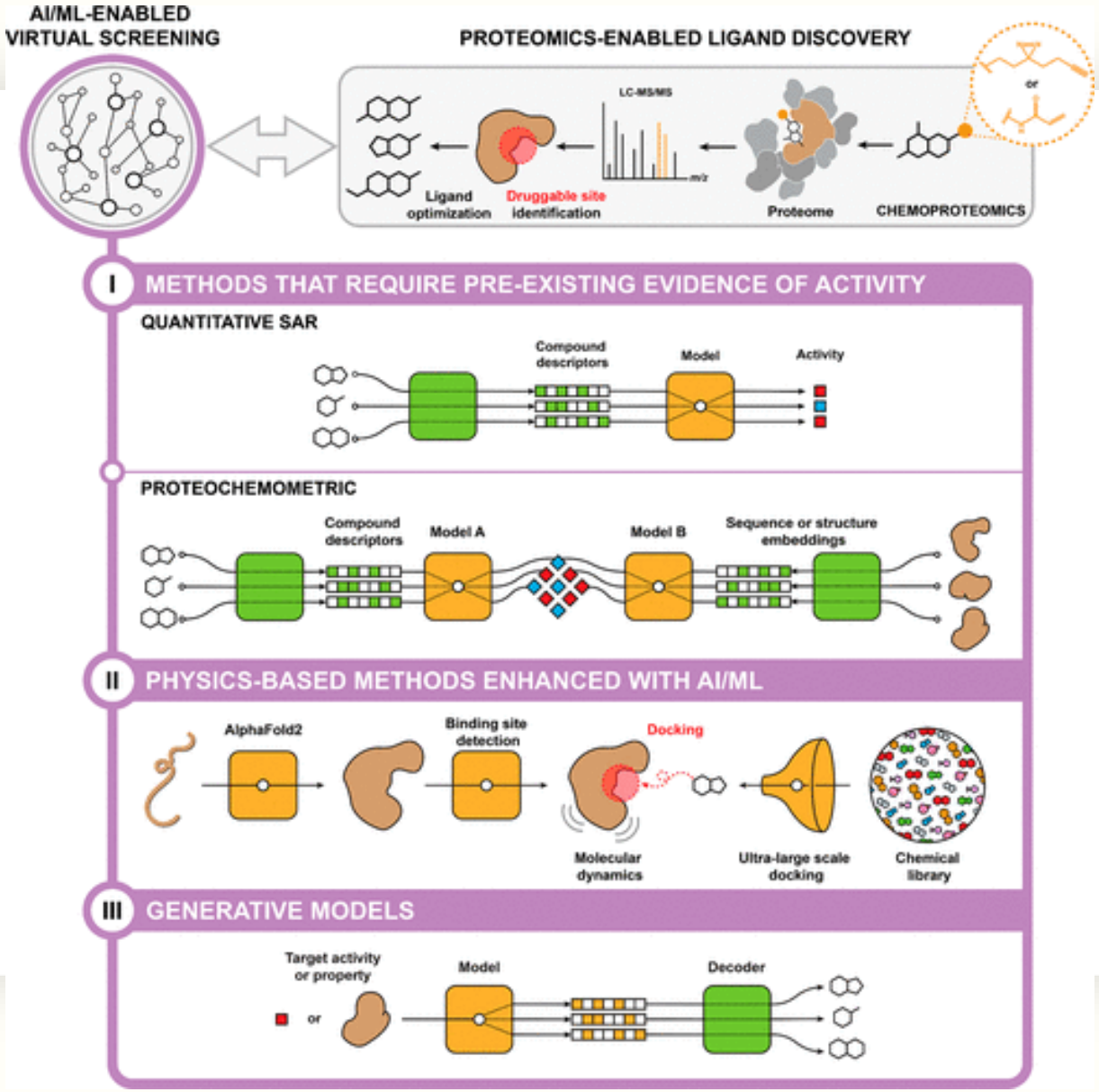
Read Online

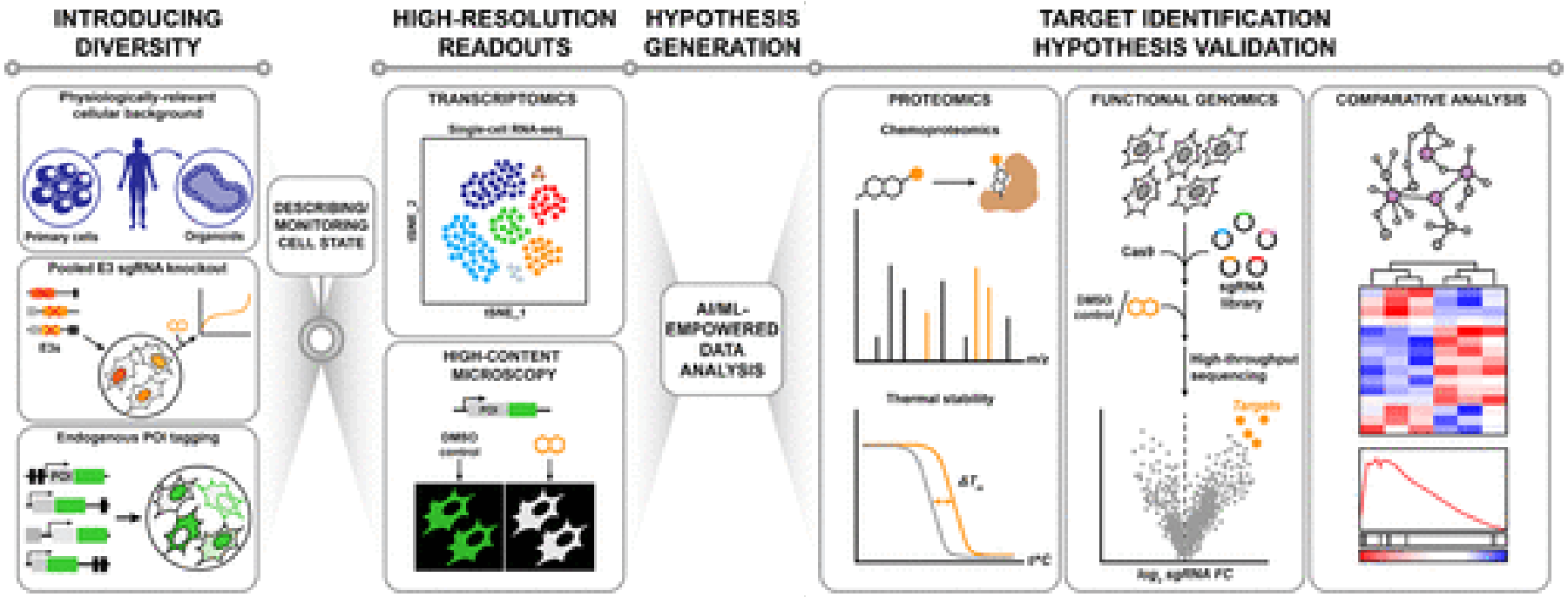
ACCESS | Metrics & More | Article Recommendations

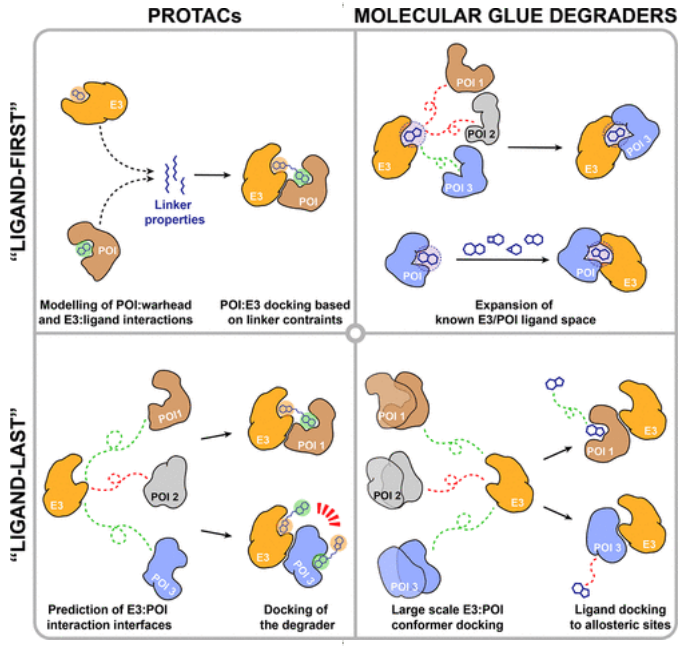
ABSTRACT: Only around 20% of the human proteome is considered to be druggable with small-molecule antagonists. This leaves some of the most compelling therapeutic targets outside the reach of ligand discovery. The concept of targeted protein degradation (TPD) promises to overcome some of these limitations. In brief, TPD is dependent on small molecules that induce the proximity between a protein of interest (POI) and an E3 ubiquitin ligase, causing ubiquitination and degradation of the POI. In this perspective, we want to reflect on current challenges in the field, and discuss how advances in multiomics profiling, artificial intelligence, and machine learning (AI/ML) will be vital in overcoming them. The presented roadmap is discussed in the context of small-molecule degraders but is equally applicable for other emerging proximity-inducing modalities.

: 07:26:57 (UTC).
timately share published articles.


ADDRESSED CHALLENGES
OPEN CHALLENGES
LACK OF BINDING SITE
SELECTIVITY
DRUGGABLE DOMAIN IS NON-FUNCTIONAL
POI
E2
RBX
E3
Adaptor
Cullin
LIGAND DISCOVERY
NEW PHENOTYPIC ASSAYS
PREDICTING POI:E3 PAIRS
OCCUPANCY-DRIVEN PHARMACOLOGY
TARGETED PROTEIN DEGRADATION


AI/ML-ENABLED
VIRTUAL SCREENING
PROTEOMICS-ENABLED LIGAND DISCOVERY
LC-MS/MS
m/z
or
Ligand
optimization
Druggable site
identification
Proteome
CHEMOPROTEOMICS
I
METHODS THAT REQUIRE PRE-EXISTING EVIDENCE OF ACTIVITY
QUANTITATIVE SAR
Compound
descriptors
Model
Activity
PROTEOCHEMOMETRIC
Compound
descriptors
Model A
Model B
Sequence or structure
embeddings
II
PHYSICS-BASED METHODS ENHANCED WITH AI/ML
AlphaFold2
Binding site
detection
Docking
Molecular
dynamics
Ultra-large scale
docking
Chemical
library
III
GENERATIVE MODELS
Target activity
or property
or
Model
Decoder


INTRODUCING DIVERSITY
HIGH-RESOLUTION READOUTS
HYPOTHESIS GENERATION
TARGET IDENTIFICATION HYPOTHESIS VALIDATION
Physiologically-relevant cellular background
Primary cells
Organoids
Pooled E3 sgRNA knockout
E3s
Endogenous POI tagging
POI
DESCRIBING/ MONITORING CELL STATE
TRANSCRIPTOMICS
Single-cell RNA-seq
tSNE_2
tSNE_1
HIGH-CONTENT MICROSCOPY
DMSO control
AI/ML-EMPOWERED DATA ANALYSIS
PROTEOMICS
Chemoproteomics
m/z
Thermal stability
ΔTm
FC
FUNCTIONAL GENOMICS
Cas9
DMSO control
sgRNA library
High-throughput sequencing
Targets
Avg. sgRNA FC
COMPARATIVE ANALYSIS


PROTACs
MOLECULAR GLUE DEGRADERS
"LIGAND-FIRST"
"LIGAND-LAST"
E3
POI
Linker properties
Modelling of POI:warhead and E3:ligand interactions
POI:E3 docking based on linker contraints
POI 1
POI 2
POI 3
Expansion of known E3/POI ligand space
Prediction of E3:POI interaction interfaces
Docking of the degrader
Large scale E3:POI conformer docking
Ligand docking to allosteric sites

nature communications

Article

https://doi.org/10.1038/s41467-022-34807-3

DeepPROTACs is a deep learning-based targeted degradation predictor for PROTACs

Received: 18 October 2021

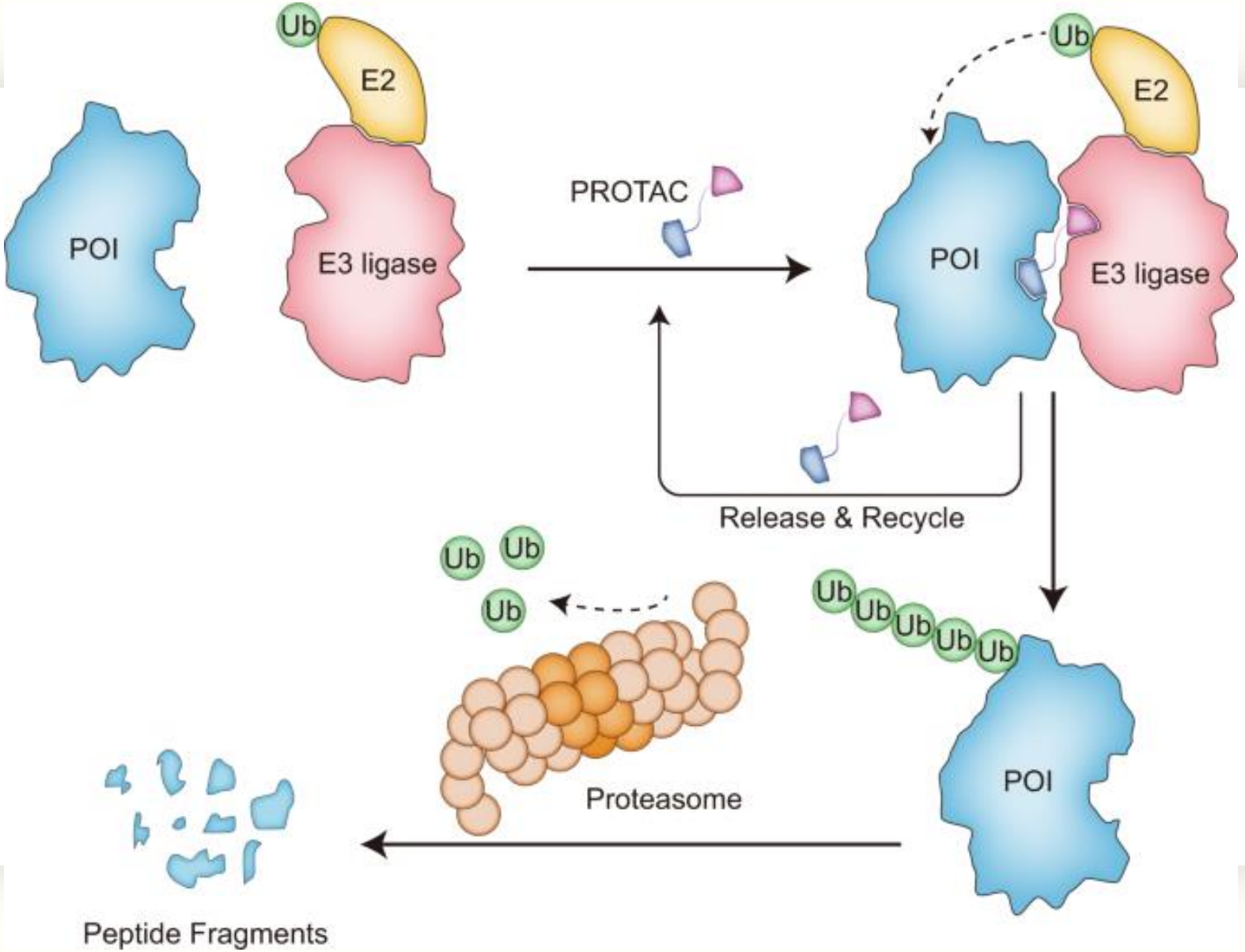
Accepted: 8 November 2022

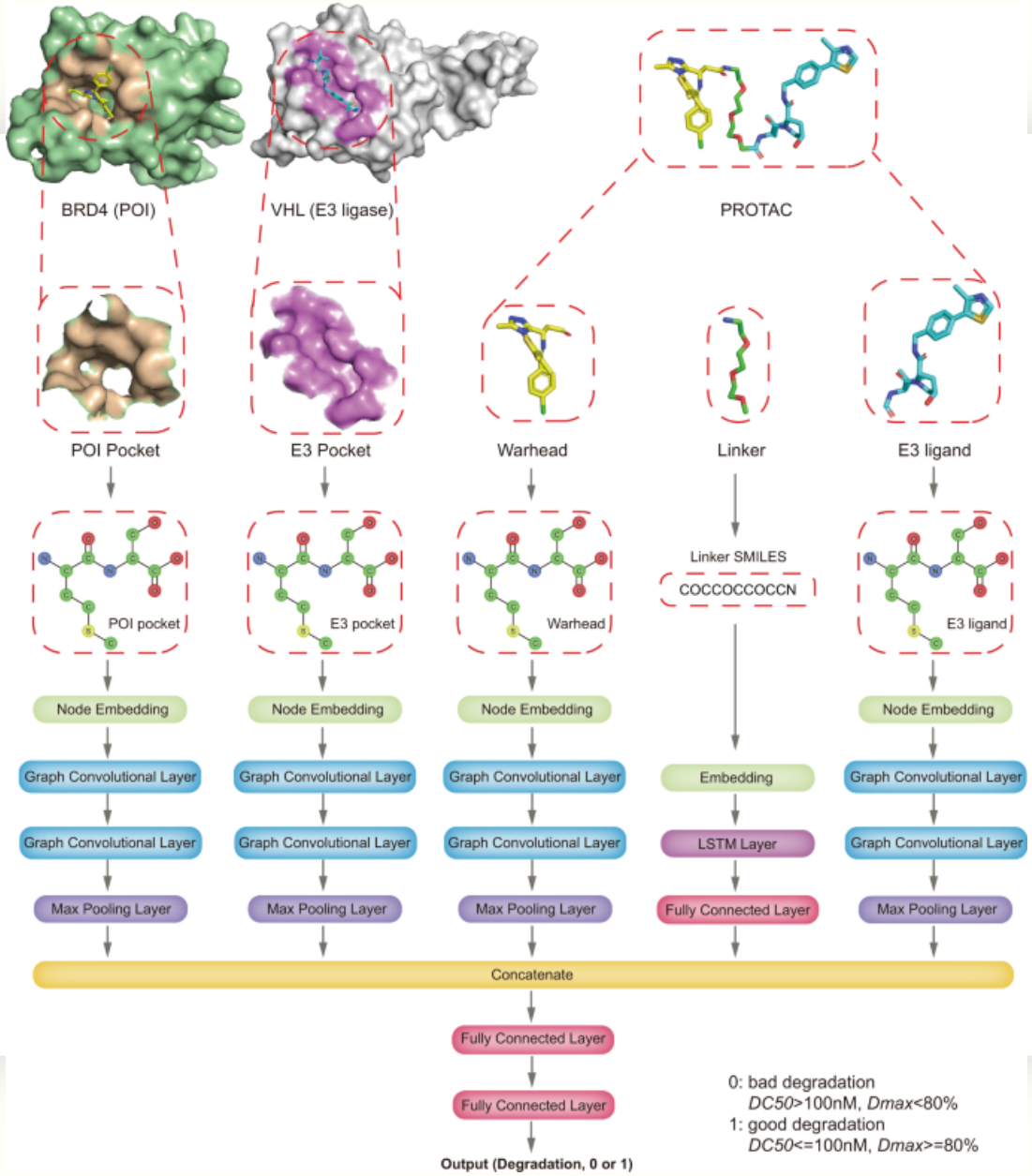
Published online: 21 November 2022

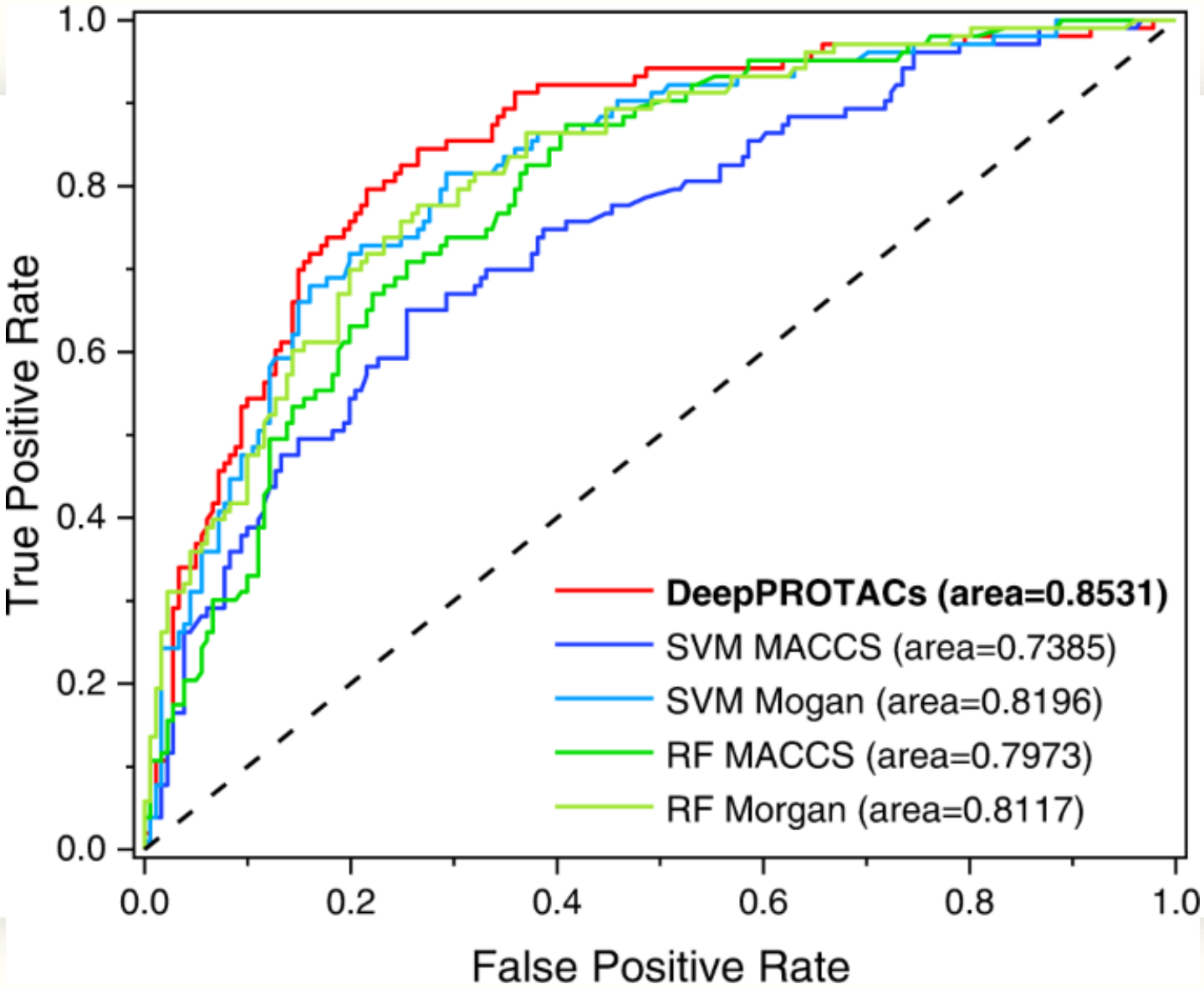
Check for updates

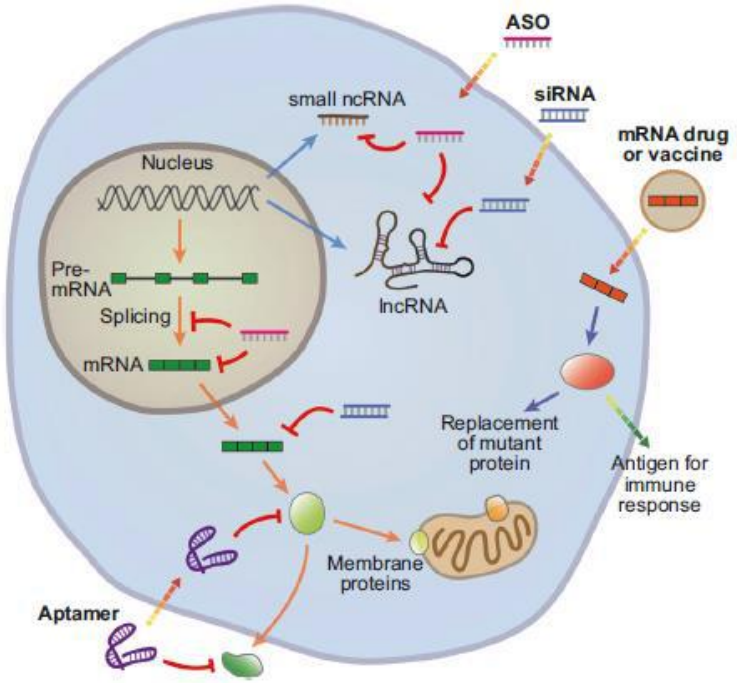
Fenglei Li[1,2,6], Qiaoyu Hu[1,6], Xianglei Zhang[1,6], Renhong Sun[3,6], Zhuanghua Liu[2,6], Sanan Wu[1], Siyuan Tian[1,4], Xinyue Ma[1,2], Zhizhuo Dai[4], Xiaobao Yang[3] ✉, Shenghua Gao[2] ✉ & Fang Bai[1,2,4,5] ✉

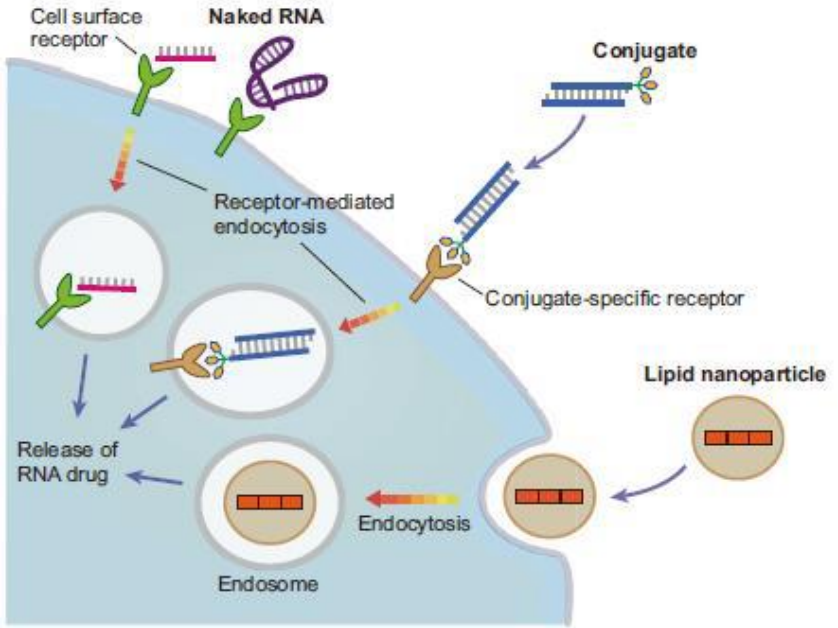
The rational design of PROTACs is difficult due to their obscure structure-activity relationship. This study introduces a deep neural network model - DeepPROTACs to help design potent PROTACs molecules. It can predict the degradation capacity of a proposed PROTAC molecule based on structures of given target protein and E3 ligase. The experimental dataset is mainly collected from PROTAC-DB and appropriately labeled according to the DC_{50} and *Dmax* values. In the model of DeepPROTACs, the ligands as well as the ligand binding


Ub
E2
POI
E3 ligase
PROTAC
Ub
E2
POI
E3 ligase
Release & Recycle
Ub
Ub
Ub
Ub
Ub
Ub
Ub
Ub
Ub
POI
Proteasome
Peptide Fragments


BRD4 (POI)
VHL (E3 ligase)
PROTAC
POI Pocket
E3 Pocket
Warhead
Linker
E3 ligand
POI pocket
E3 pocket
Warhead
Linker SMILES
COCCOCCOCCN
E3 ligand
Node Embedding
Graph Convolutional Layer
Graph Convolutional Layer
Max Pooling Layer
Embedding
LSTM Layer
Fully Connected Layer
Concatenate
Fully Connected Layer
Fully Connected Layer
Output (Degradation, 0 or 1)
0: bad degradation
DC50>100nM, Dmax<80%
1: good degradation
DC50<=100nM, Dmax>=80%


True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
DeepPROTACs (area=0.8531)
SVM MACCS (area=0.7385)
SVM Mogan (area=0.8196)
RF MACCS (area=0.7973)
RF Morgan (area=0.8117)





* Source: xxxx, 2022.0x.xx

frontiers in Bioengineering and Biotechnology

REVIEW
published: 18 March 2021
doi: 10.3389/fbioe.2021.628137

Check for updates

The Limitless Future of RNA Therapeutics

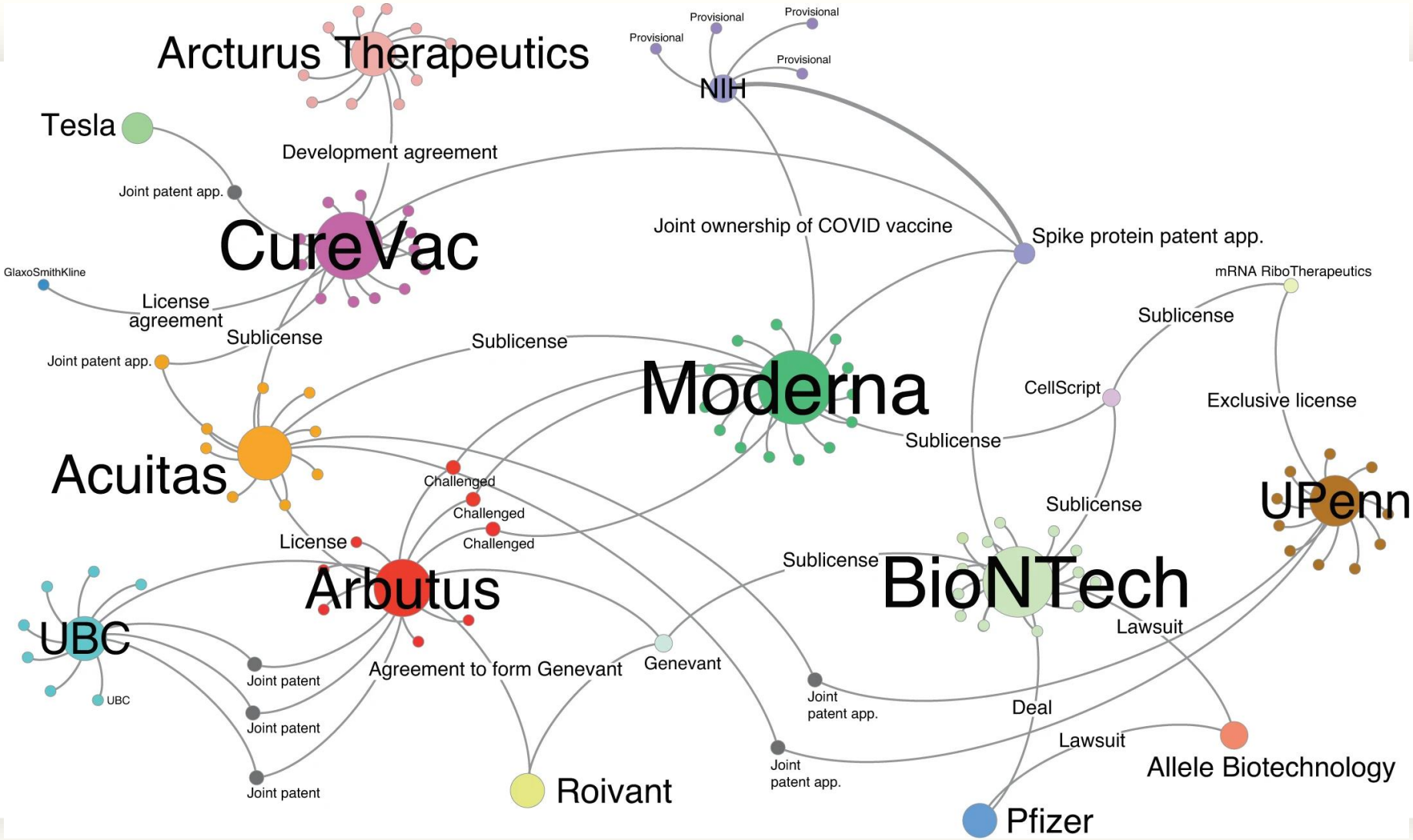
*Tulsi Ram Damase[1], Roman Sukhovershin[1], Christian Boada[2], Francesca Taraballi[3,4], Roderic I. Pettigrew[2] and John P. Cooke[1]**

[1] RNA Therapeutics Program, Department of Cardiovascular Sciences, Houston Methodist Research Institute, Houston, TX, United States, [2] Colleges of Medicine, Engineering, Texas A&M University and Houston Methodist Hospital, Houston, TX, United States, [3] Center for Musculoskeletal Regeneration, Houston Methodist Research Institute, Houston, TX, United States, [4] Department of Orthopedics and Sports Medicine, Houston Methodist Hospital, Houston, TX, United States

TABLE 1 | RNA therapeutics approved for clinical use or undergoing clinical trials.

Drug	Type of RNA	Company	Route of delivery	Condition/Disease	Status
Nusinersen (Spinraza)	ASO	Ionis	Intrathecal	Spinal muscular atrophy	FDA approval in 2016
Eteplirsen (Exondys 51)	ASO	Sarepta	Intravenous	Duchenne muscular dystrophy	FDA approval in 2016
Inotersen (Tegsedi)	ASO	Ionis	Subcutaneous	Familial amyloid polyneuropathy	FDA approval in 2018
Volanesorsen (Waylivra)	ASO	Ionis	Subcutaneous	Familial chylomicronemia syndrome	EU approval in 2019
Patisiran (Onpattro)	siRNA	Alnylam	Intravenous	Polyneuropathy	FDA approval in 2018
Givosiran (Givlaari)	siRNA	Alnylam	Subcutaneous	Acute hepatic porphyria	FDA approval in 2019
Cobomarsen (MRG-106)	miRNA	miRagen (Viridian)	Intravenous/subcutaneous	Blood cancers	Phase II
Remlarsen (MRG-201)	miRNA	miRagen (Viridian)	Intradermal	Keloids	Phase II
MRG-110	miRNA	miRagen (Viridian)	Intradermal	Tissue Repair	Phase I
Pegaptanib (Macugen)	Aptamer(RNA)	Bausch + Lomb	Intravitreal	Macular Degeneration	FDA approval in 2014
Emapticap pegol (NOX-E36)	Aptamer(RNA)	NOXXON	Intravenous/Subcutaneous	Diabetic nephropathy, lung and pancreatic cancer	Phase II
Olaptesed pegol (NOX-A12)	Aptamer(RNA)	NOXXON	Intravenous	Brain cancer	Phase I/II
BNT162b2	mRNA	BioNTech and Pfizer	Intramuscular	COVID-19	FDA authorization for emergency use in 2020
mRNA-1273	mRNA	Moderna	Intramuscular	COVID-19	FDA authorization for emergency use in 2020
CVnCoV	mRNA	CureVac	Intramuscular	COVID-19	Phase III
AZD8601	mRNA	Moderna/Astrazeneca	Epicardial	Ischemic heart disease	Phase II
mRNA-1647	mRNA	Moderna	Intramuscular	Cytomegalovirus infection	Phase II
P-BCMA-101	mRNA	Poseida	Intravenous	Multiple myeloma	Phase II
mRNA-4157	mRNA	Moderna	Intramuscular	Cancer	Phase II
mRNA-3704	mRNA	Moderna	Intravenous	Methylmalonic aciduria	Phase I/II
MRT5005	mRNA	Translate Bio	Inhalation	Cystic Fibrosis	Phase I/II
mRNA-2416	mRNA	Moderna	Intratumoral	Solid tumors/ lymphoma/advanced ovarian carcinoma	Phase I/II
BNT131 (SAR441000)	mRNA	BioNTech/ Sanofi/Genmab	Intratumoral	Advanced melanoma	Phase I/II
Descartes-08	mRNA	Cartesian	Intravenous	Generalized myasthenia gravis	Phase I/II
BNT122	mRNA	BioNTech	Intravenous	Solid tumors	Phase I/II
mRNA-2752	mRNA	Moderna	Intratumoral	Solid tumors	Phase I
MEDI1191	mRNA	Moderna	Intratumoral	Solid tumors	Phase I
mRNA-1944	mRNA	Moderna	Intravenous	Chikungunya infection	Phase I
CV8102	mRNA	CureVac	Intratumoral	Solid tumors	Phase I
ARCT-810	mRNA	Arcturus	Intravenously	Urea disorder	Phase I
CV7202	mRNA	CureVac	Intramuscular	Rabies	Phase I
mRNA-1893	mRNA	Moderna	Intramuscular	Zika	Phase I
CV9202	mRNA	CureVac	Intradermal	Non-small cell lung cancer	Phase I
mRNA-5671	mRNA	Moderna	Intravenous	Cancer	Phase I
BNT111	mRNA	BioNTech	Intravenous	Advanced Melanoma	Phase I

The table summarizes information available at ClinicalTrials.gov (https://clinicaltrials.gov/) as of January 26, 2021. ASO, antisense oligonucleotide; siRNA, small interfering RNA; miRNA, microRNA; RNA, ribonucleic acid; mRNA, messenger RNA.



Gaviria, M., Kilic, B. A network analysis of COVID-19 mRNA vaccine patents. *Nat Biotechnol* **39**, 546–548 (2021). https://doi.org/10.1038/s41587-021-00912-9

Computational and Structural Biotechnology Journal 20 (2022) 5680–5689

ELSEVIER

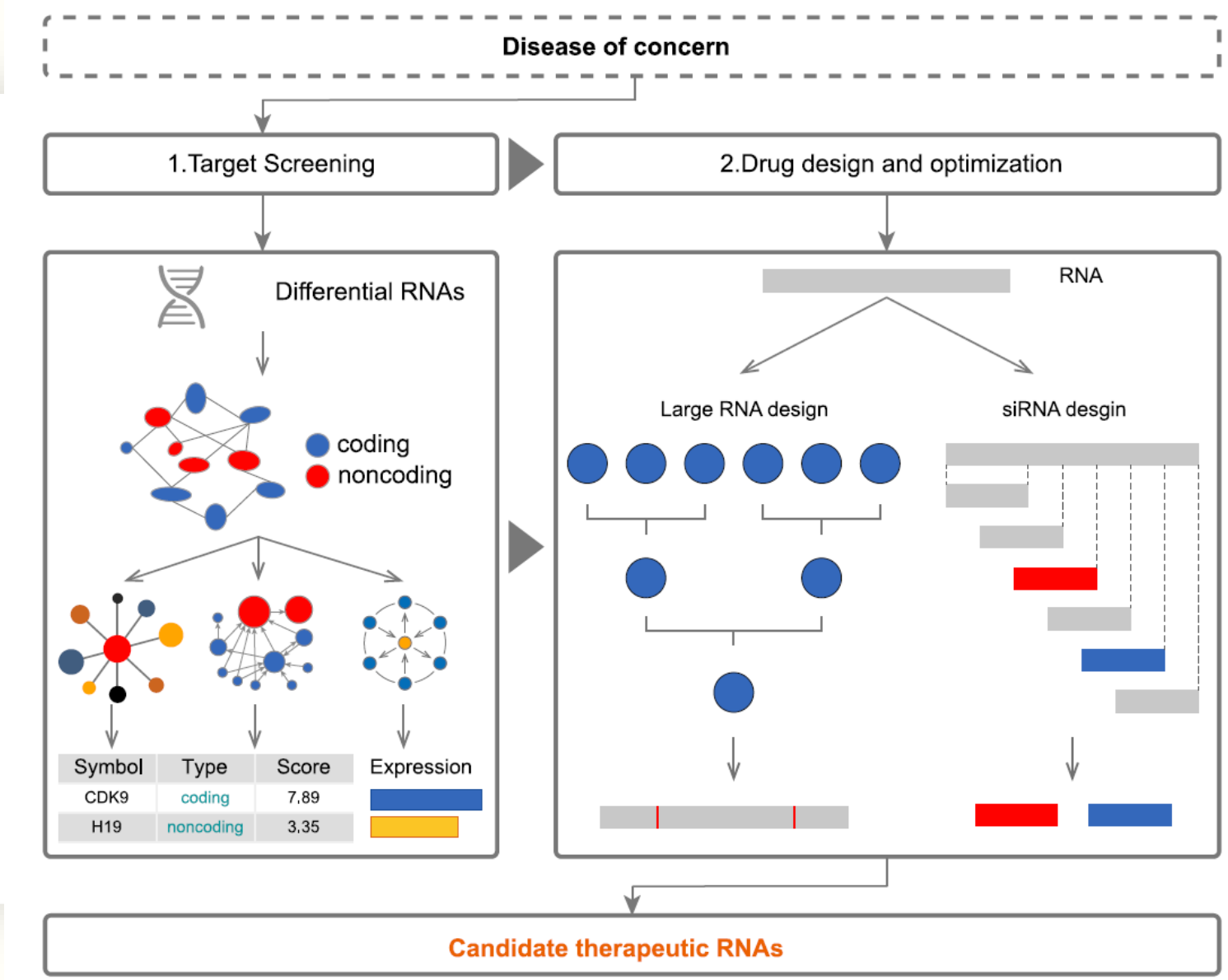
COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

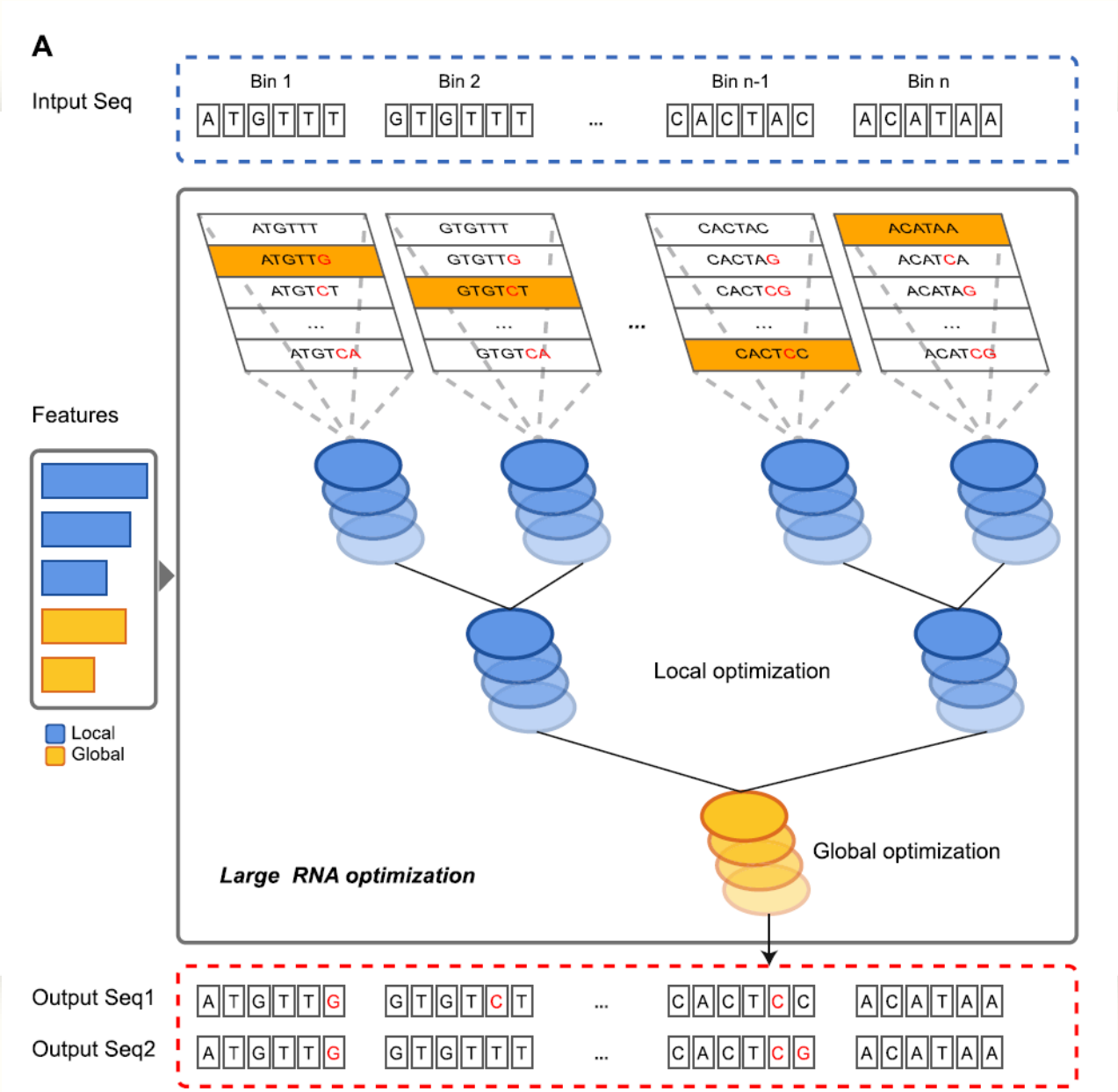
journal homepage: www.elsevier.com/locate/csbj

TREAT: Therapeutic RNAs exploration inspired by artificial intelligence technology

Yufan Luo [a,b,1], Liu Liu [a,1], Zihao He [a,1], Shanshan Zhang [c], Peipei Huo [c], Zhihao Wang [c], Qin Jiaxin [a], Lianhe Zhao [a], Yang Wu [a], Dongdong Zhang [e], Dechao Bu [a,d,*], Runsheng Chen [e,f,*], Yi Zhao [a,*]

[a] *Research Center for Ubiquitous Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China*
[b] *University of Chinese Academy of Sciences, Beijing 100049, China*
[c] *Luoyang Zhongke Information Industry Research Institute, Luoyang, China*
[d] *Hwa Mei Hospital, University of Chinese Academy of Sciences, China*
[e] *Key Laboratory of RNA Biology, Center for Big Data Research in Health, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China*
[f] *Shenzhen Institute of Nucleic Acid Drug Research, Shenzhen Bay Laboratory Pingshan Translational Medicine Center, Shenzhen 510800, China*


Disease of concern
1.Target Screening
2.Drug design and optimization
Differential RNAs
coding
noncoding
Symbol
Type
Score
Expression
CDK9
coding
7.89
H19
noncoding
3.35
RNA
Large RNA design
siRNA desgin
Candidate therapeutic RNAs


A
Intput Seq
Bin 1
Bin 2
Bin n-1
Bin n
ATGTTT GTGTTT CACTAC ACATAA
ATGTTG GTGTTG CACTAG ACATCA
ATGTCT GTGTCT CACTCG ACATAG
ATGTCA GTGTCA CACTCC ACATCG
Features
Local
Global
Local optimization
Global optimization
Large RNA optimization
Output Seq1
Output Seq2

nature machine intelligence

Article

https://doi.org/10.1038/s42256-022-00571-8

Deep learning models for predicting RNA degradation via dual crowdsourcing

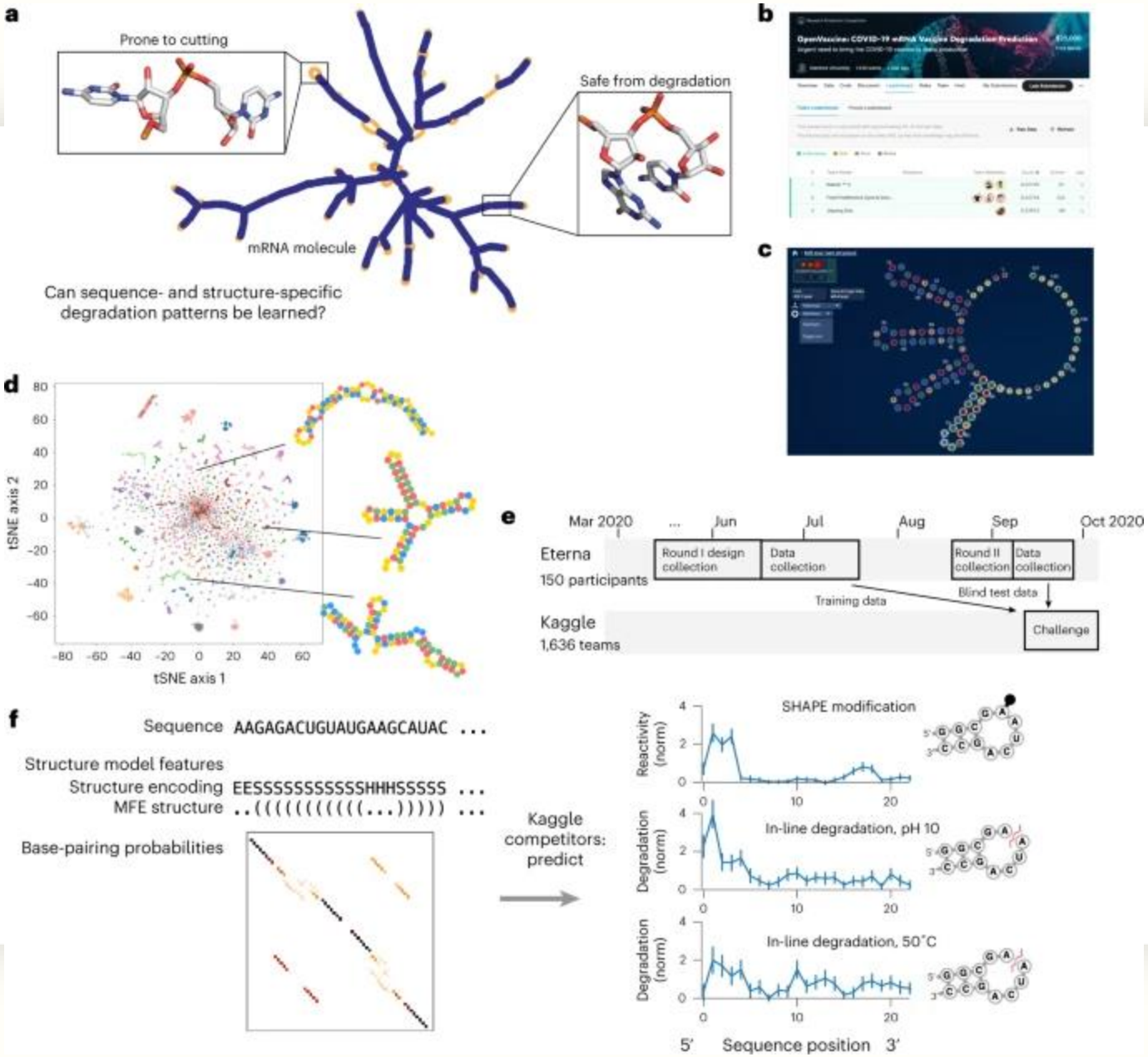
Received: 14 October 2021

Accepted: 21 October 2022

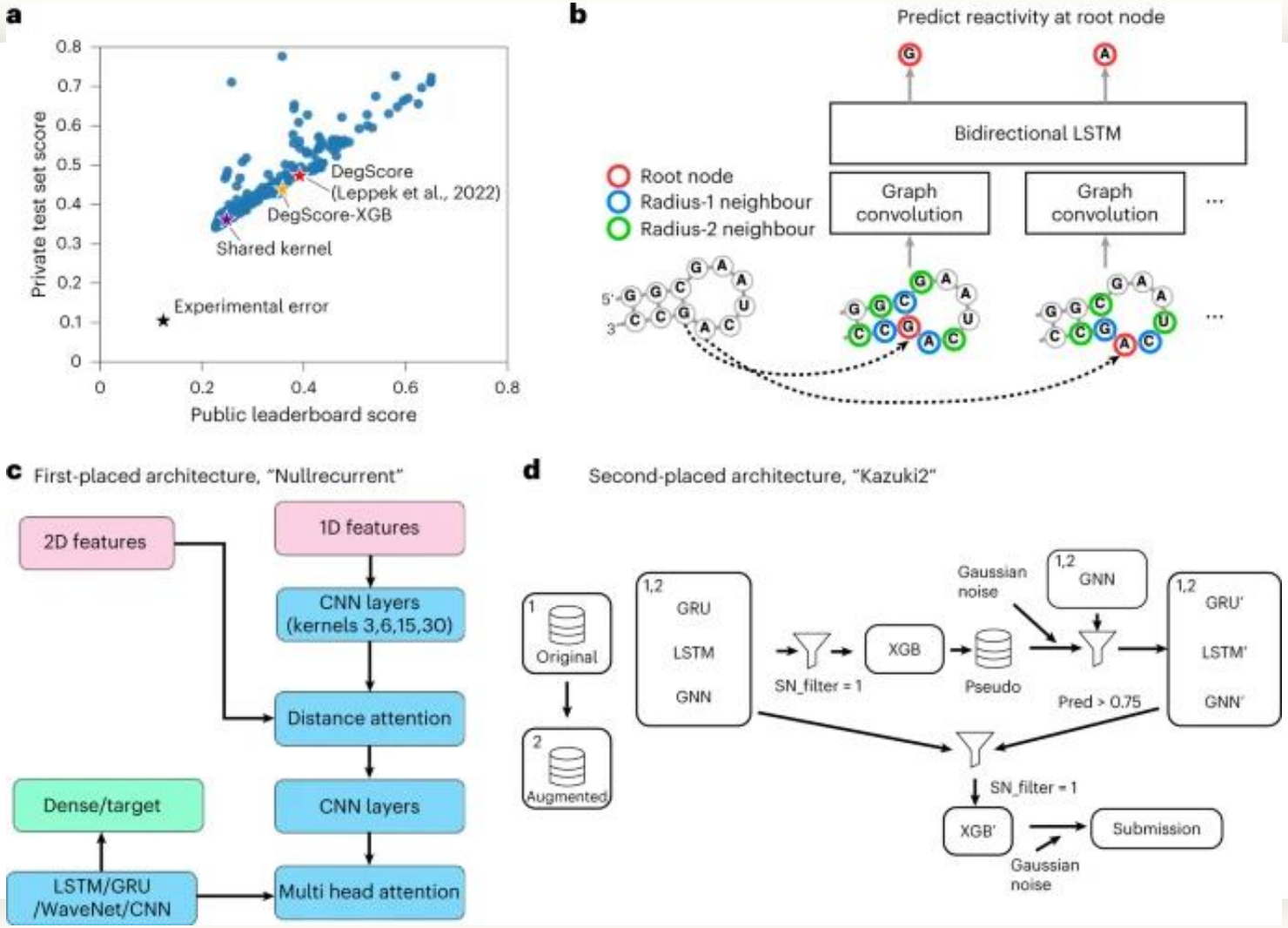
Published online: 14 December 2022

Check for updates

Hannah K. Wayment-Steele[1,2,26], Wipapat Kladwang[2,3,26], Andrew M. Watkins[2,3,4,26], Do Soon Kim[2,3,26], Bojan Tunguz[3,5,26], Walter Reade[6], Maggie Demkin[6], Jonathan Romano[2,3,7], Roger Wellington-Oguri[2], John J. Nicol[2], Jiayang Gao[8], Kazuki Onodera[9], Kazuki Fujikawa[10], Hanfei Mao[11], Gilles Vandewiele[12], Michele Tinti[13], Bram Steenwinckel[12], Takuya Ito[14], Taiga Noumi[15], Shujun He[16], Keiichiro Ishi[17], Youhan Lee[18,19], Fatih Öztürk[20], King Yuen Chiu[21], Emin Öztürk[22], Karim Amer[23], Mohamed Fares[23,24], Eterna Participants* & Rhiju Das[2,3,25]



* Source: xxxx, 2022.0x.xx



* Source: xxxx, 2022.0x.xx

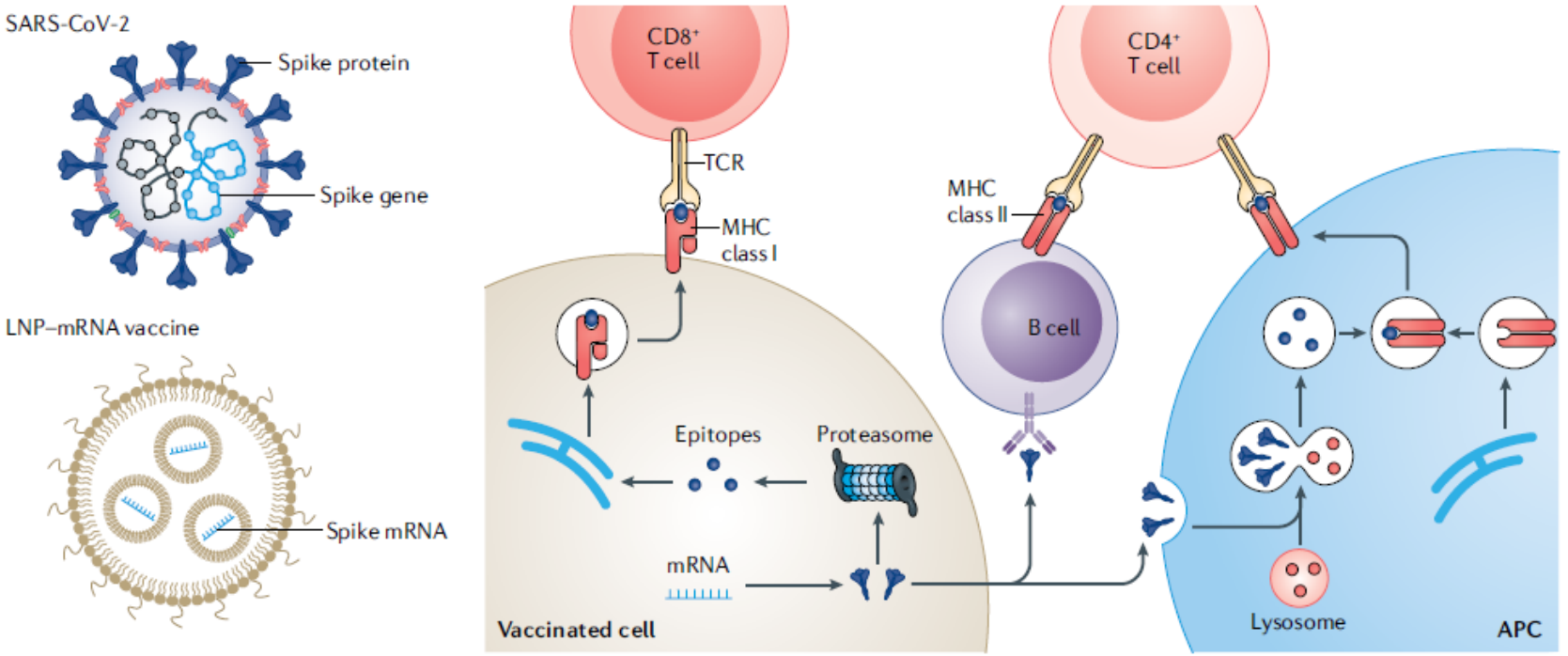
REVIEWS

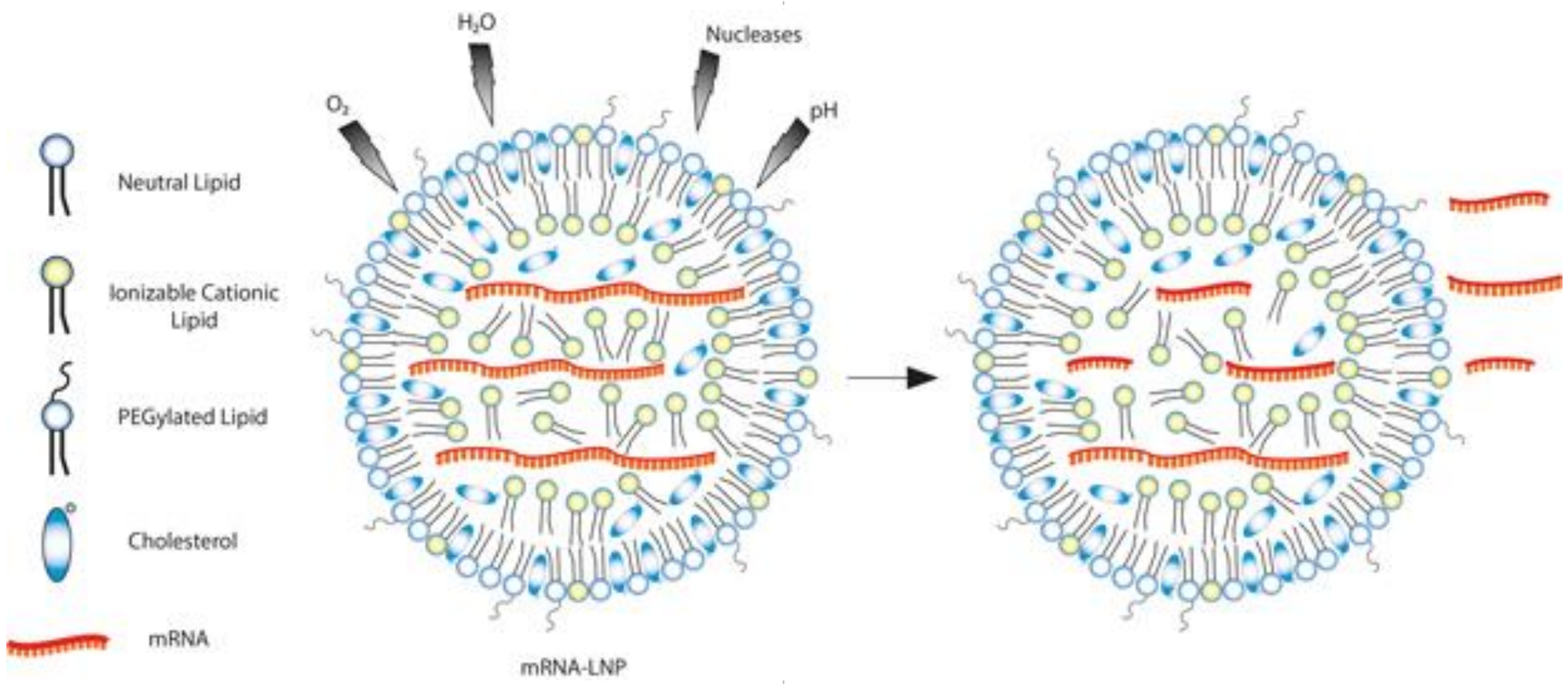
Check for updates

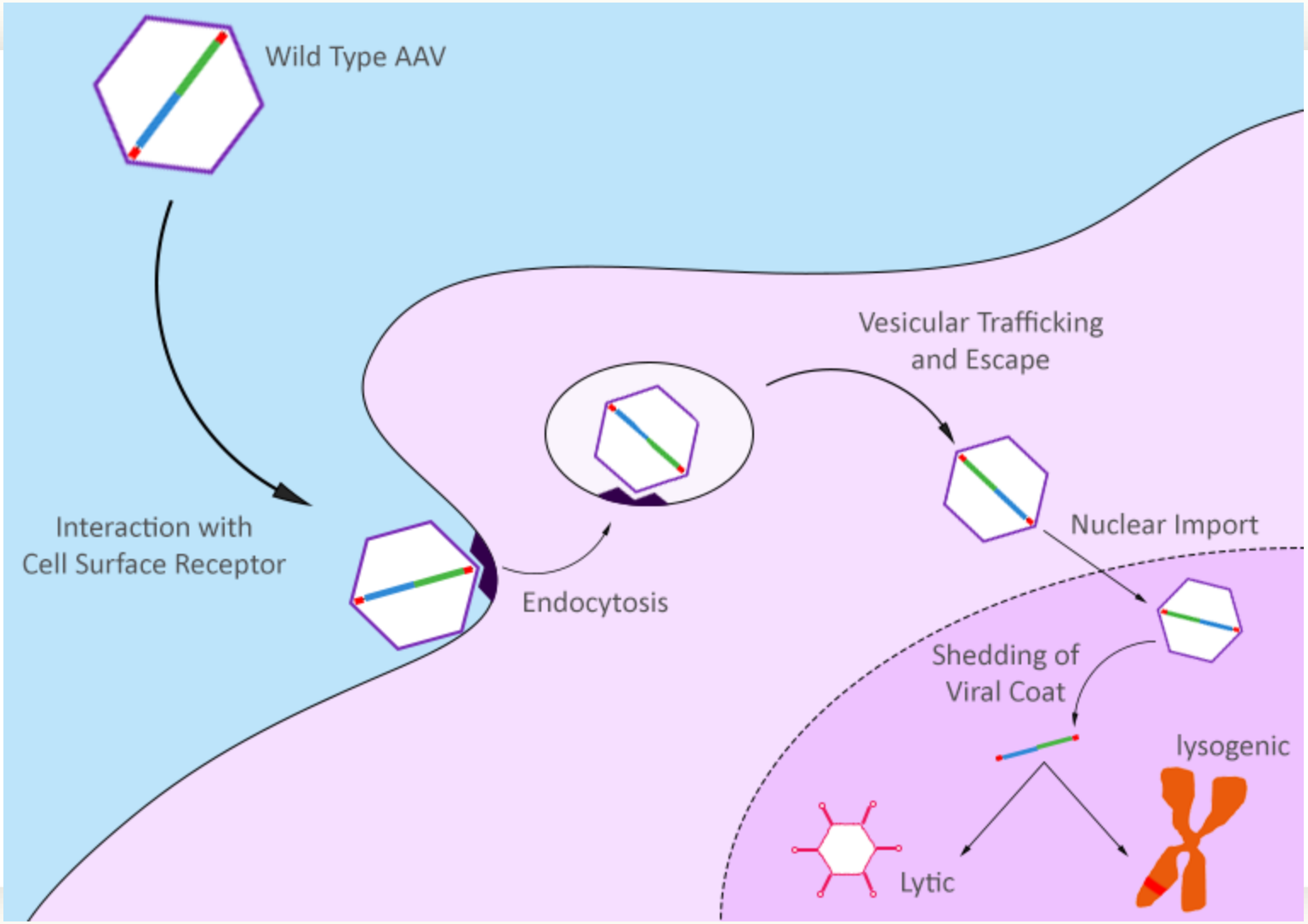
Lipid nanoparticles for mRNA delivery

Xucheng Hou[1], *Tal Zaks*[2]✉, *Robert Langer*[3,4]✉ *and Yizhou Dong*[1]✉

Abstract | Messenger RNA (mRNA) has emerged as a new category of therapeutic agent to prevent and treat various diseases. To function in vivo, mRNA requires safe, effective and stable delivery systems that protect the nucleic acid from degradation and that allow cellular uptake and mRNA release. Lipid nanoparticles have successfully entered the clinic for the delivery of mRNA; in particular, lipid nanoparticle–mRNA vaccines are now in clinical use against coronavirus disease 2019 (COVID-19), which marks a milestone for mRNA therapeutics. In this Review, we discuss the design of lipid nanoparticles for mRNA delivery and examine physiological barriers and possible administration routes for lipid nanoparticle–mRNA systems. We then consider key points for the clinical translation of lipid nanoparticle–mRNA formulations, including good manufacturing practice, stability, storage and safety, and highlight preclinical and clinical studies of lipid nanoparticle–mRNA therapeutics for infectious diseases, cancer and genetic disorders. Finally, we give an outlook to future possibilities and remaining challenges for this promising technology.


SARS-CoV-2
Spike protein
Spike gene
LNP–mRNA vaccine
Spike mRNA
CD8+ T cell
TCR
MHC class I
Epitopes
Proteasome
mRNA
Vaccinated cell
CD4+ T cell
MHC class II
B cell
Lysosome
APC


Neutral Lipid
Ionizable Cationic Lipid
PEGylated Lipid
Cholesterol
mRNA
H_2O
Nucleases
O_2
pH
mRNA-LNP


Wild Type AAV
Interaction with
Cell Surface Receptor
Endocytosis
Vesicular Trafficking
and Escape
Nuclear Import
Shedding of
Viral Coat
lysogenic
Lytic

Deep-learning models for lipid nanoparticle-based drug delivery

Philip J Harrison*,[1], Håkan Wieslander[2], Alan Sabirsh[3], Johan Karlsson[4], Victor Malmsjö[1], Andreas Hellander[2], Carolina Wählby[2] & Ola Spjuth[1]

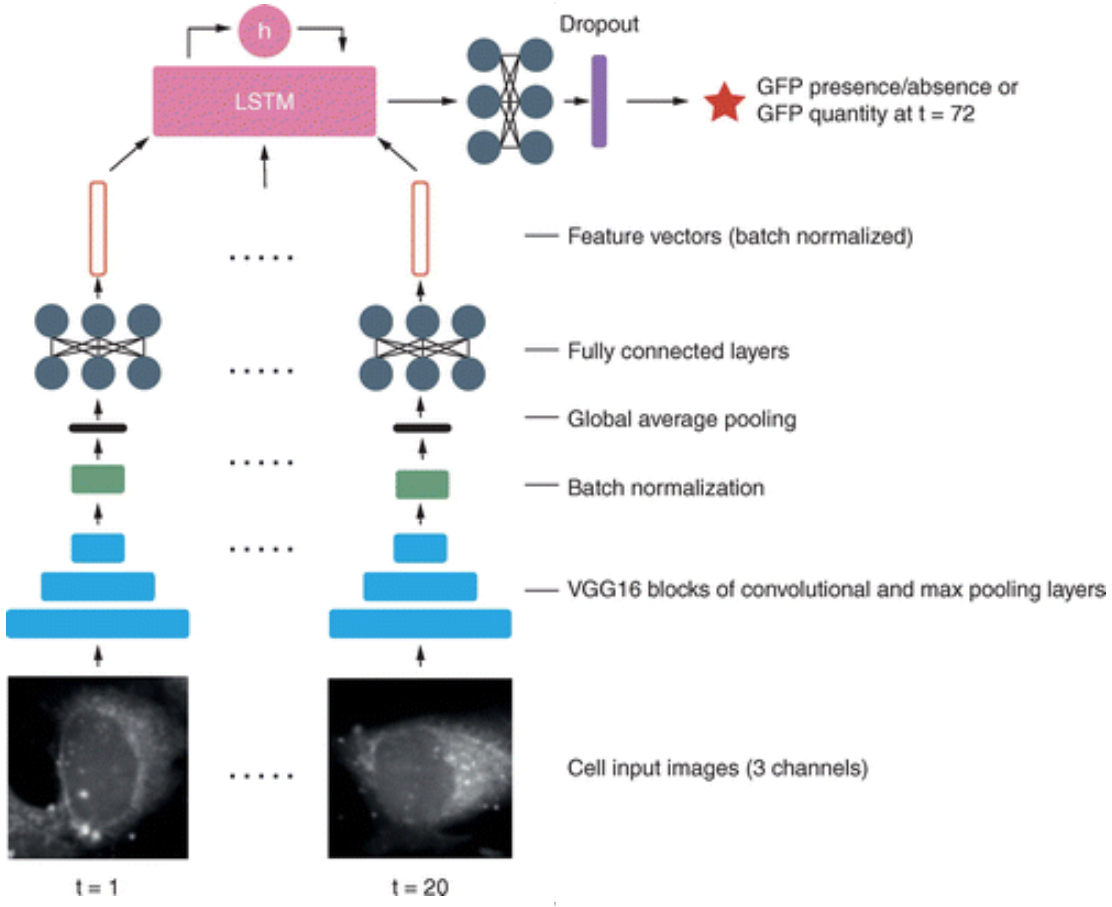
[1]Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden

[2]Department of Information Technology, Uppsala University, Uppsala, Sweden

[3]Advanced Drug Delivery, Pharmaceutical Sciences, Research and Development, AstraZeneca, Gothenburg, Sweden

[4]Quantitative Biology, Discovery Sciences, Research and Development, AstraZeneca, Gothenburg, Sweden

*Author for correspondence: philip.harrison@farmbio.uu.se

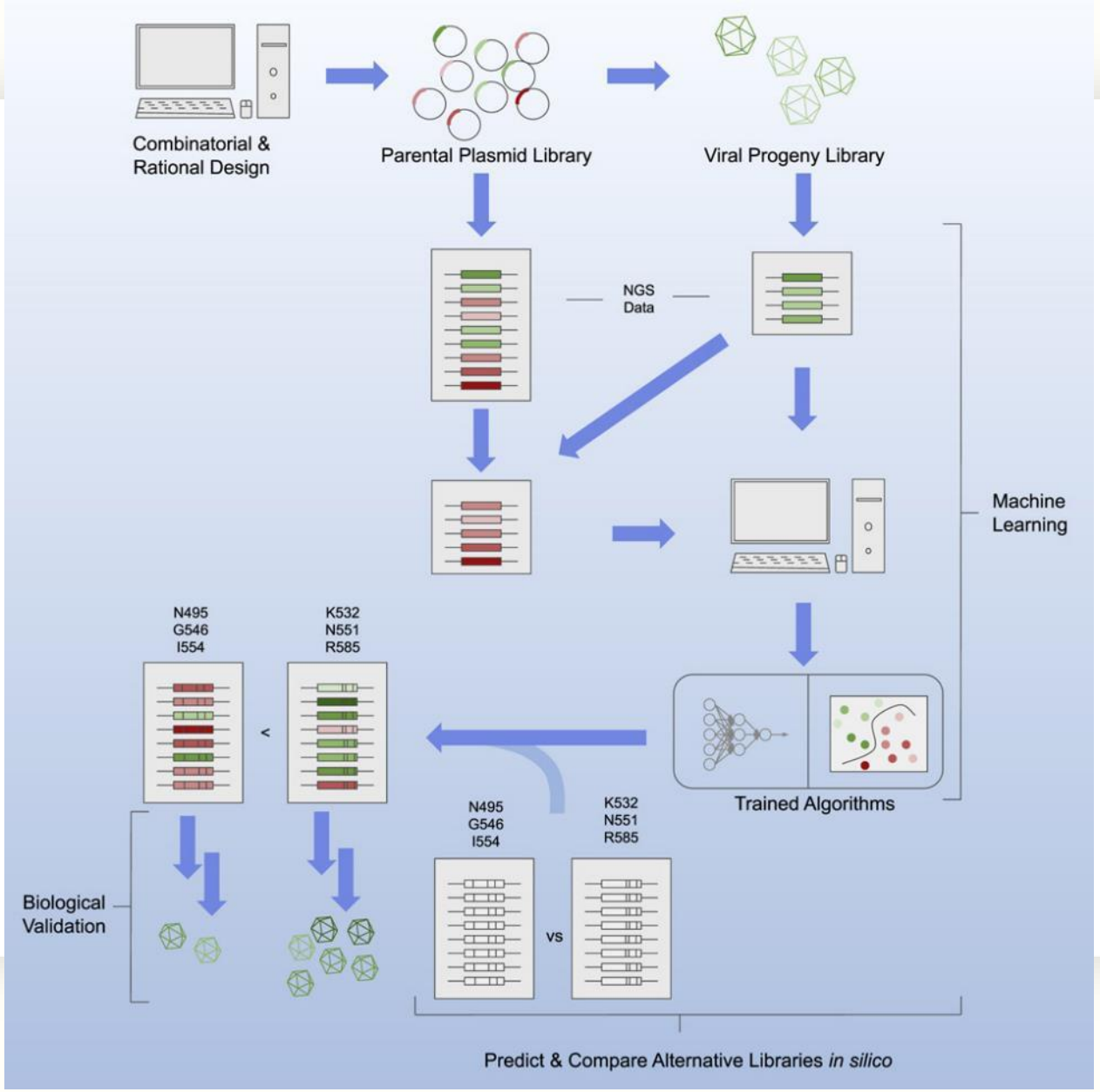

h
LSTM
Dropout
GFP presence/absence or
GFP quantity at t = 72
Feature vectors (batch normalized)
Fully connected layers
Global average pooling
Batch normalization
VGG16 blocks of convolutional and max pooling layers
Cell input images (3 channels)
t = 1
t = 20

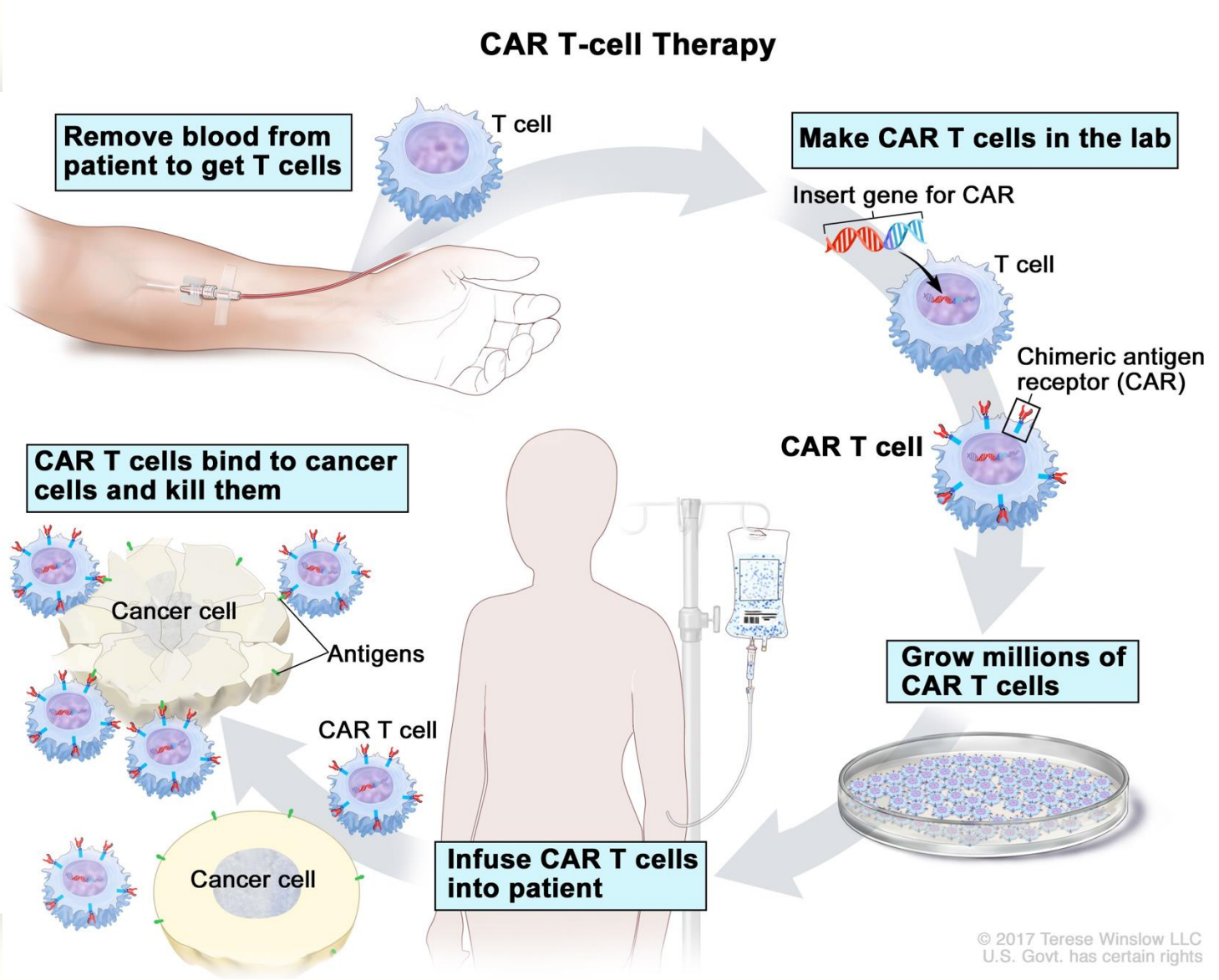
Molecular Therapy
Methods & Clinical Development
Original Article

Applying machine learning to predict viral assembly for adeno-associated virus capsid libraries

Andrew D. Marques,[1] Michael Kummer,[2] Oleksandr Kondratov,[1] Arunava Banerjee,[2] Oleksandr Moskalenko,[3] and Sergei Zolotukhin[1]

[1]Department of Pediatrics, Division of Cellular and Molecular Therapy, University of Florida, Gainesville, FL 32608, USA; [2]Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32603, USA; [3]University of Florida Research Computing, University of Florida, Gainesville, FL 32608, USA


Combinatorial &
Rational Design
Parental Plasmid Library
Viral Progeny Library
NGS
Data
Machine
Learning
N495
G546
I554
K532
N551
R585
<
Trained Algorithms
Biological
Validation
N495
G546
I554
K532
N551
R585
vs
Predict & Compare Alternative Libraries *in silico*

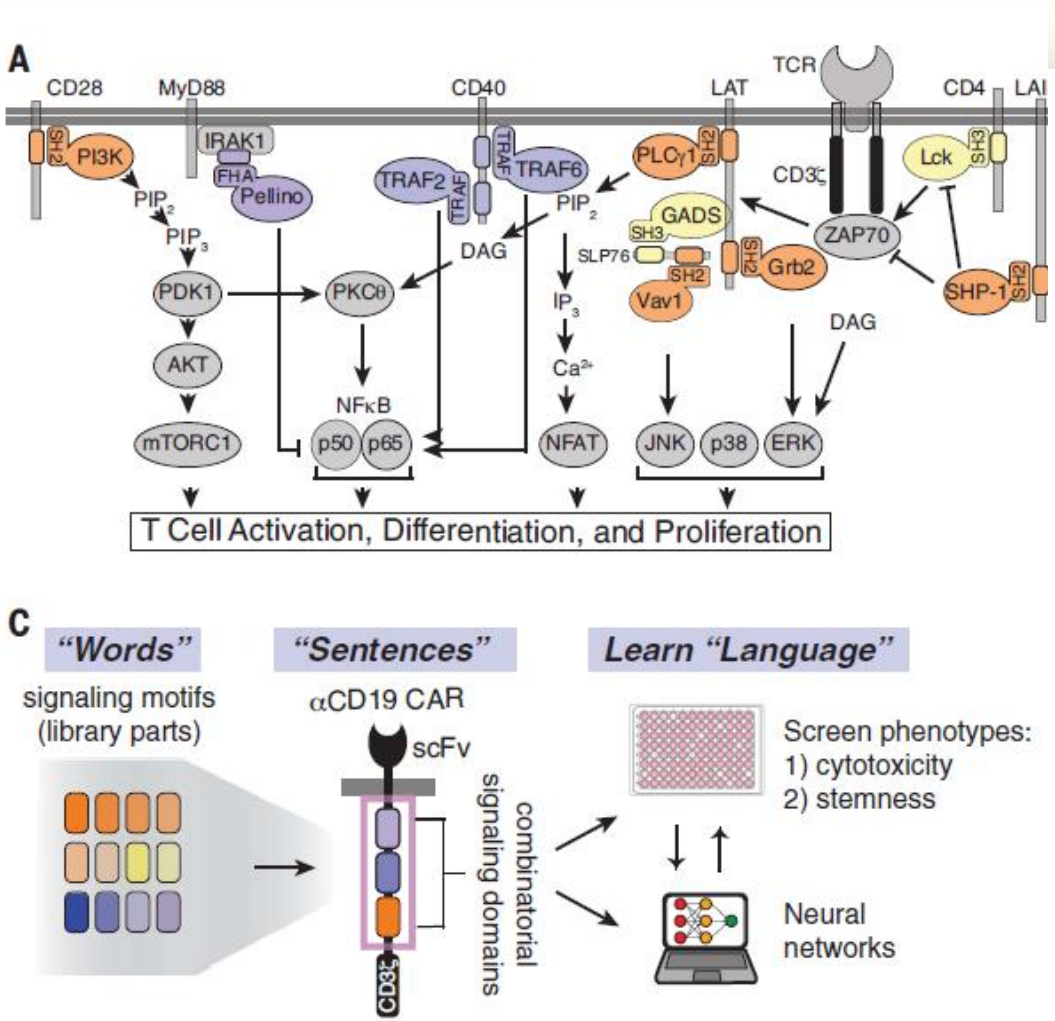

CAR T-cell Therapy
Remove blood from patient to get T cells
T cell
Make CAR T cells in the lab
Insert gene for CAR
T cell
Chimeric antigen receptor (CAR)
CAR T cell
CAR T cells bind to cancer cells and kill them
Cancer cell
Antigens
CAR T cell
Cancer cell
Infuse CAR T cells into patient
Grow millions of CAR T cells
© 2017 Terese Winslow LLC
U.S. Govt. has certain rights

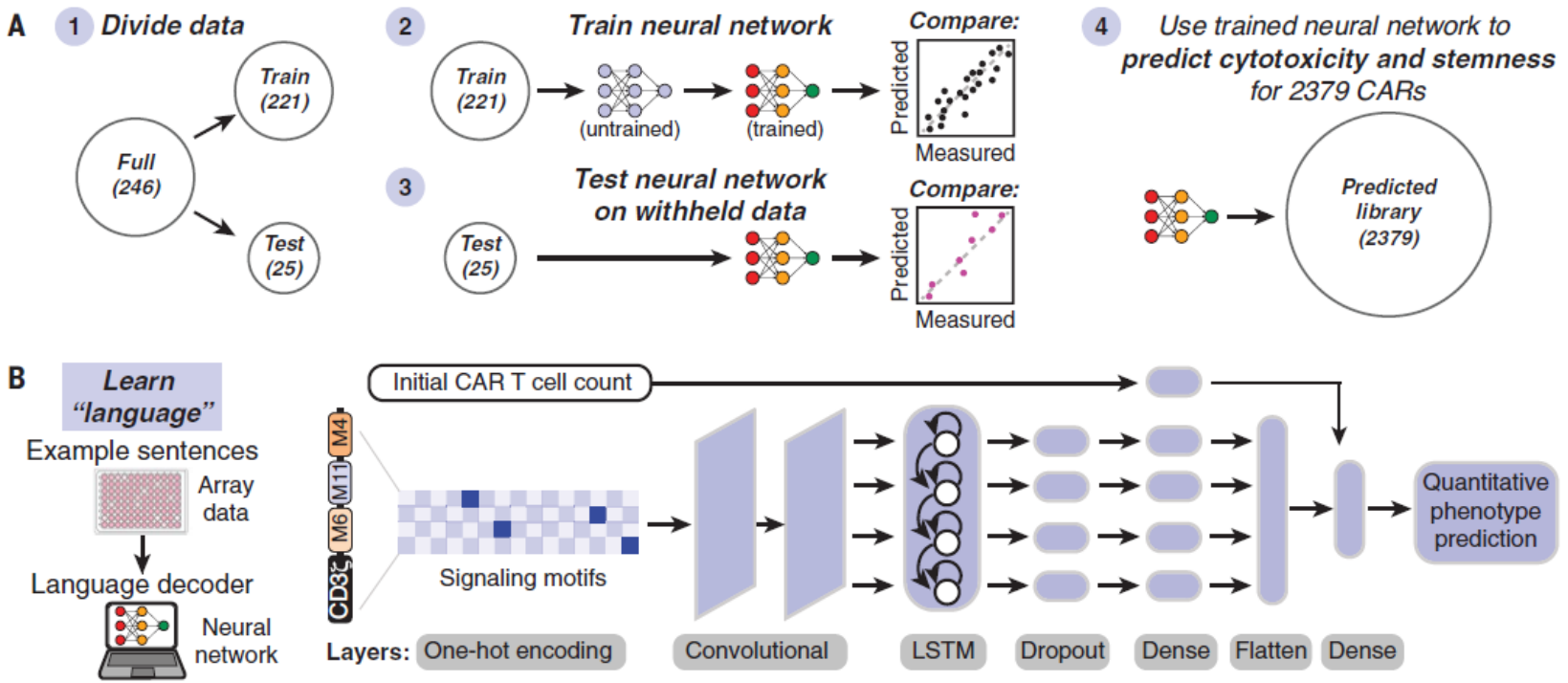
SYNTHETIC BIOLOGY

Decoding CAR T cell phenotype using combinatorial signaling motif libraries and machine learning

Kyle G. Daniels[1,2], Shangying Wang[3,4], Milos S. Simic[1,2], Hersh K. Bhargava[1,2], Sara Capponi[3,4], Yurie Tonai[1,2], Wei Yu[1,2], Simone Bianco[3,4]†*, Wendell A. Lim[1,2,4]*

Chimeric antigen receptor (CAR) costimulatory domains derived from native immune receptors steer the phenotypic output of therapeutic T cells. We constructed a library of CARs containing ~2300 synthetic costimulatory domains, built from combinations of 13 signaling motifs. These CARs promoted diverse human T cell fates, which were sensitive to motif combinations and configurations. Neural networks trained to decode the combinatorial grammar of CAR signaling motifs allowed extraction of key design rules. For example, non-native combinations of motifs that bind tumor necrosis factor receptor–associated factors (TRAFs) and phospholipase C gamma 1 (PLCγ1) enhanced cytotoxicity and stemness associated with effective tumor killing. Thus, libraries built from minimal building blocks of signaling, combined with machine learning, can efficiently guide engineering of receptors with desired phenotypes.


A
CD28
MyD88
CD40
LAT
TCR
CD4
LAI
SH2
PI3K
IRAK1
FHA
Pellino
TRAF2
TRAF
TRAF6
PLCγ1
CD3ζ
Lck
SH3
PIP2
PIP3
DAG
GADS
SLP76
ZAP70
Grb2
Vav1
SHP-1
PDK1
PKCθ
IP3
Ca2+
AKT
NFκB
mTORC1
p50
p65
NFAT
JNK
p38
ERK
T Cell Activation, Differentiation, and Proliferation
C
"Words"
"Sentences"
Learn "Language"
signaling motifs
(library parts)
αCD19 CAR
scFv
CD3ζ
combinatorial
signaling domains
Screen phenotypes:
1) cytotoxicity
2) stemness
Neural
networks


A
1 Divide data
Full (246)
Train (221)
Test (25)
2
Train neural network
Train (221)
(untrained)
(trained)
Compare:
Predicted
Measured
3
Test neural network on withheld data
Test (25)
Compare:
Predicted
Measured
4
Use trained neural network to predict cytotoxicity and stemness for 2379 CARs
Predicted library (2379)
B
Learn "language"
Example sentences
Array data
Language decoder
Neural network
Initial CAR T cell count
CD3ζ
M6
M11
M4
Signaling motifs
Quantitative phenotype prediction
Layers:
One-hot encoding
Convolutional
LSTM
Dropout
Dense
Flatten
Dense

nature reviews drug discovery

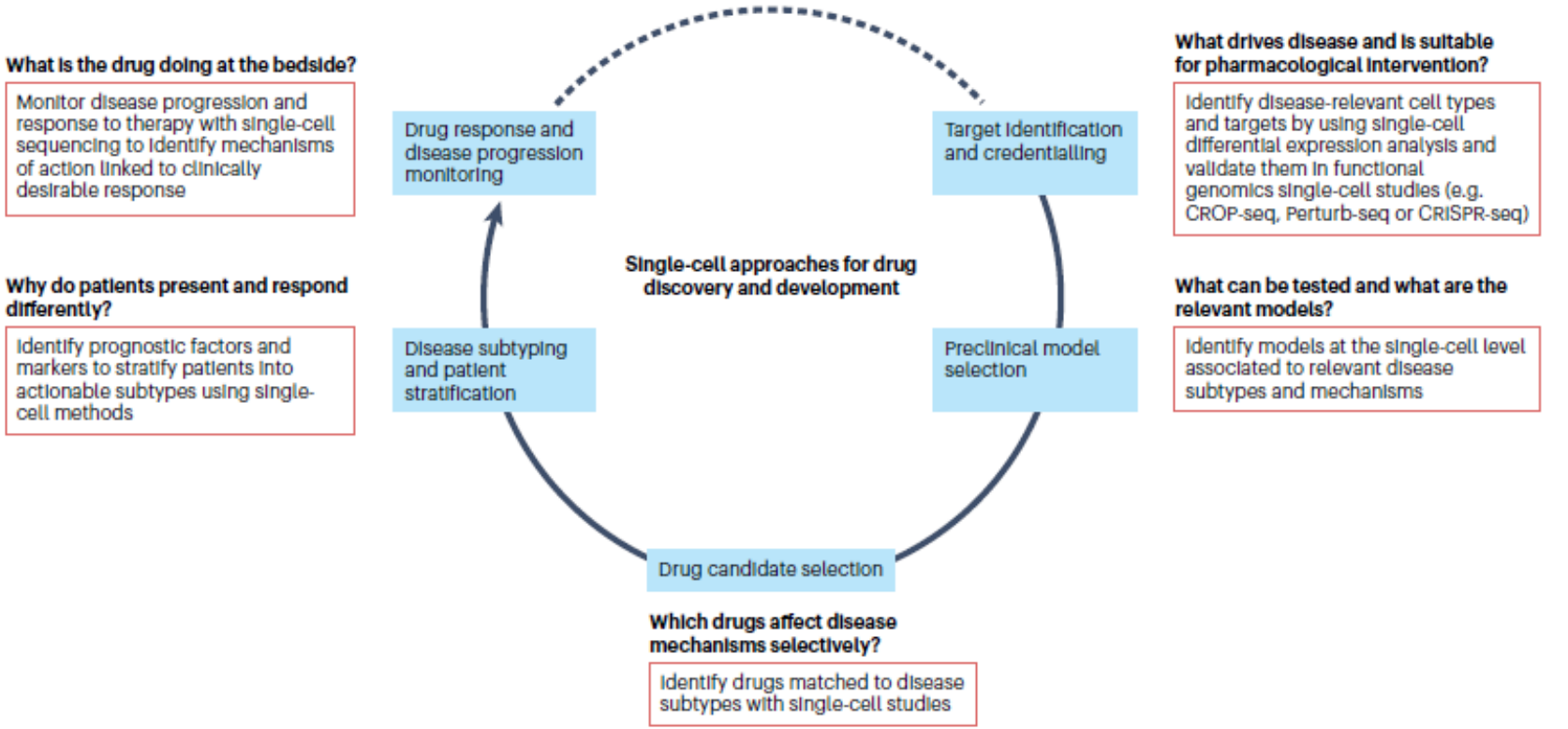
https://doi.org/10.1038/s41573-023-00688-4

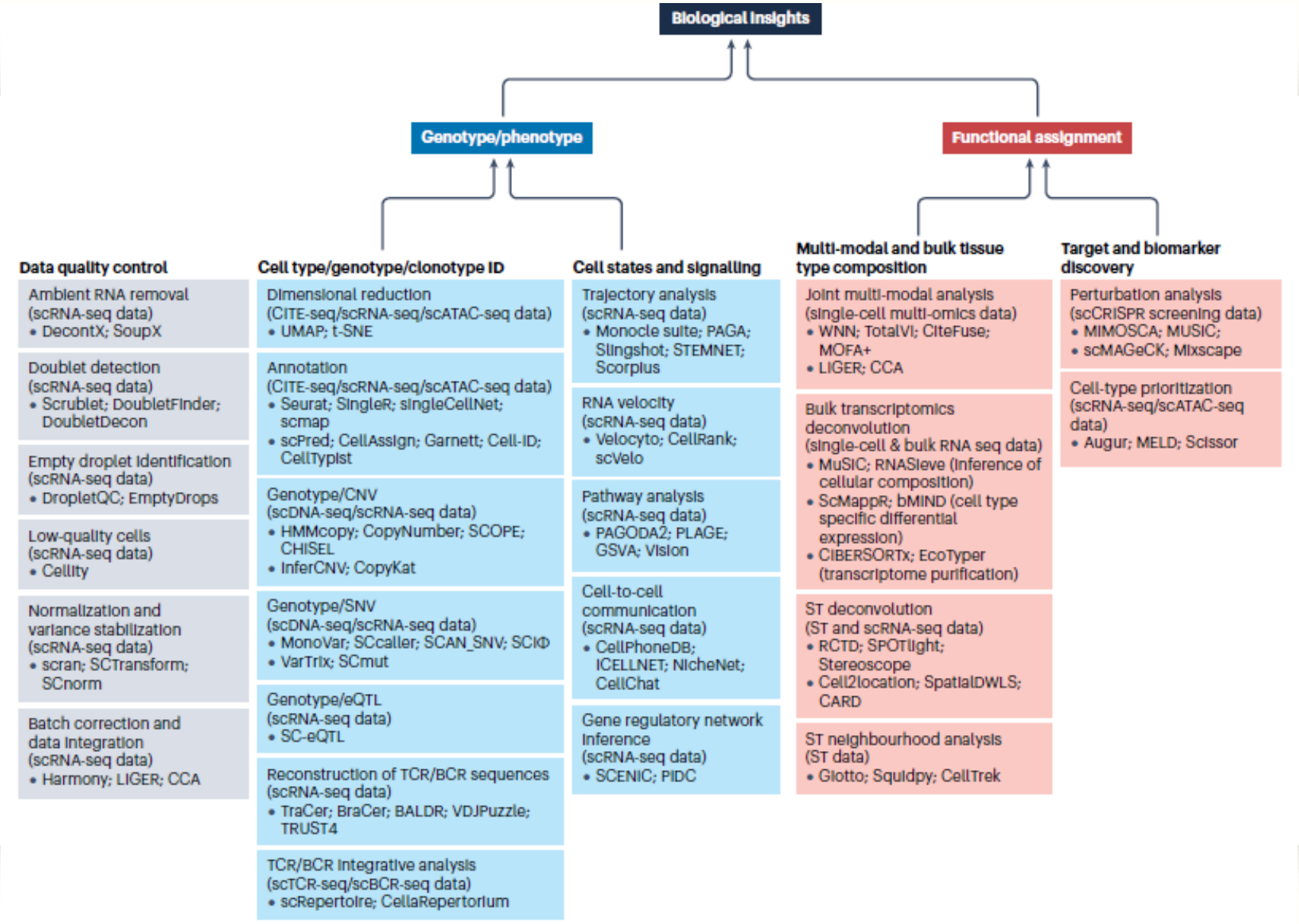
Review article

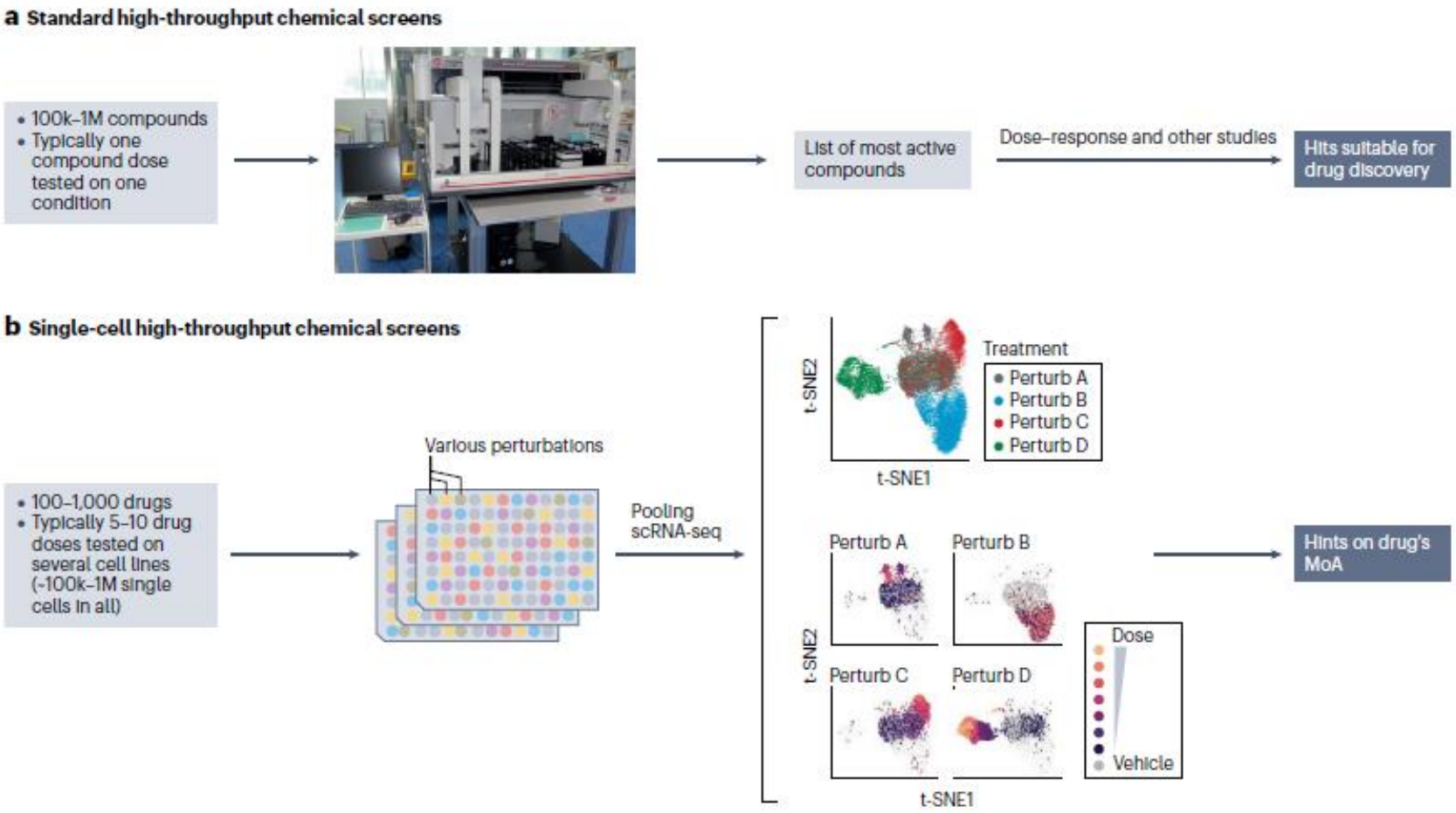
Check for updates

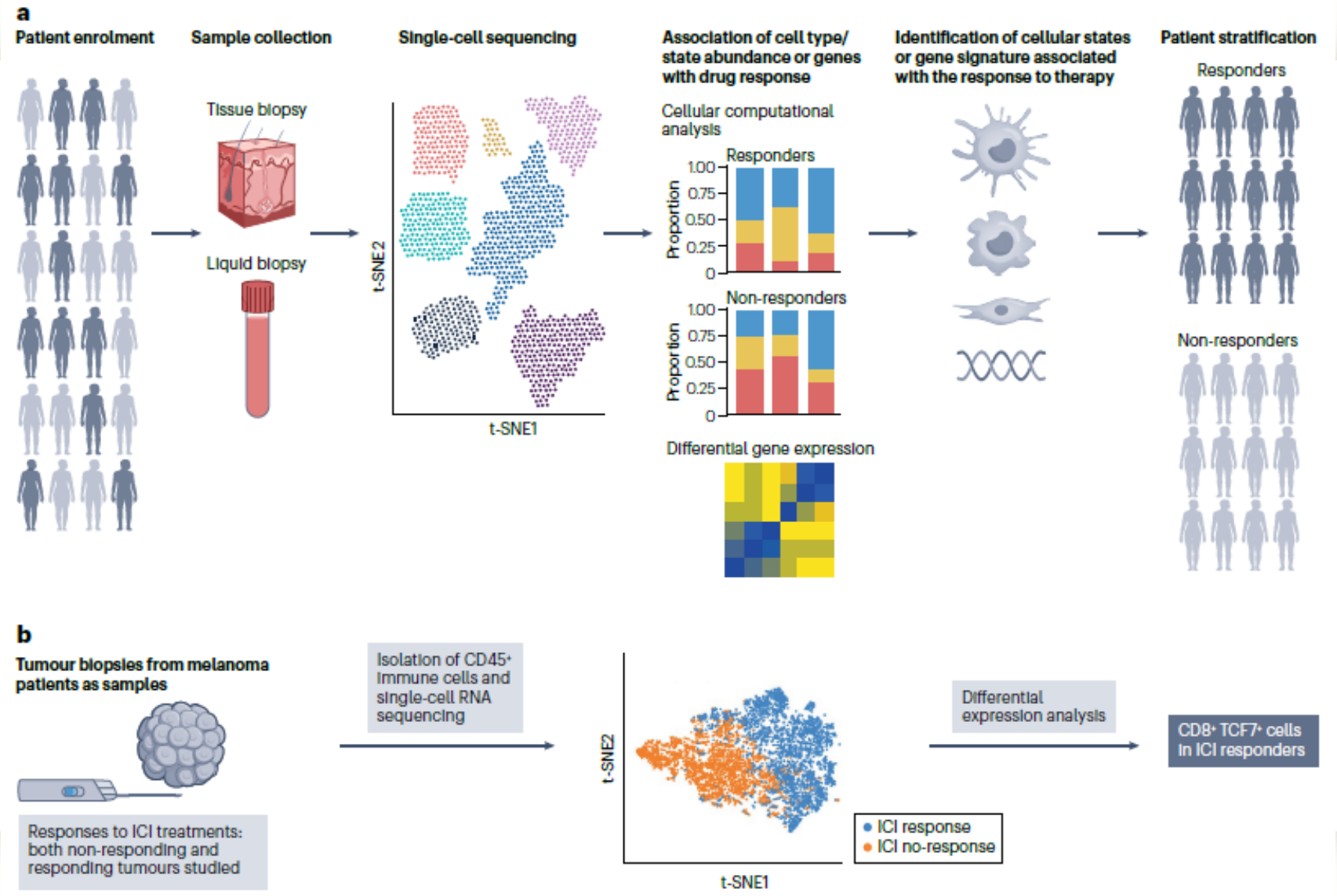
Applications of single-cell RNA sequencing in drug discovery and development

Bram Van de Sande[1,15], Joon Sang Lee[2,15], Euphemia Mutasa-Gottgens[3,15] ✉, Bart Naughton[4], Wendi Bacon[3,5], Jonathan Manning[3], Yong Wang[6], Jack Pollard[7], Melissa Mendez[8], Jon Hill[9], Namit Kumar[10], Xiaohong Cao[11], Xiao Chen[12], Mugdha Khaladkar[13], Ji Wen[14], Andrew Leach[3] & Edgardo Ferran[3]


What is the drug doing at the bedside?
Monitor disease progression and response to therapy with single-cell sequencing to identify mechanisms of action linked to clinically desirable response
Why do patients present and respond differently?
Identify prognostic factors and markers to stratify patients into actionable subtypes using single-cell methods
Drug response and disease progression monitoring
Disease subtyping and patient stratification
Single-cell approaches for drug discovery and development
Target identification and credentialling
Preclinical model selection
Drug candidate selection
What drives disease and is suitable for pharmacological intervention?
Identify disease-relevant cell types and targets by using single-cell differential expression analysis and validate them in functional genomics single-cell studies (e.g. CROP-seq, Perturb-seq or CRISPR-seq)
What can be tested and what are the relevant models?
Identify models at the single-cell level associated to relevant disease subtypes and mechanisms
Which drugs affect disease mechanisms selectively?
Identify drugs matched to disease subtypes with single-cell studies


Biological insights
Genotype/phenotype
Functional assignment
Data quality control
Ambient RNA removal (scRNA-seq data)
• DecontX; SoupX
Doublet detection (scRNA-seq data)
• Scrublet; DoubletFinder; DoubletDecon
Empty droplet identification (scRNA-seq data)
• DropletQC; EmptyDrops
Low-quality cells (scRNA-seq data)
• Cellity
Normalization and variance stabilization (scRNA-seq data)
• scran; SCTransform; SCnorm
Batch correction and data integration (scRNA-seq data)
• Harmony; LIGER; CCA
Cell type/genotype/clonotype ID
Dimensional reduction (CITE-seq/scRNA-seq/scATAC-seq data)
• UMAP; t-SNE
Annotation (CITE-seq/scRNA-seq/scATAC-seq data)
• Seurat; SingleR; singleCellNet; scmap
• scPred; CellAssign; Garnett; Cell-ID; CellTypist
Genotype/CNV (scDNA-seq/scRNA-seq data)
• HMMcopy; CopyNumber; SCOPE; CHISEL
• InferCNV; CopyKat
Genotype/SNV (scDNA-seq/scRNA-seq data)
• MonoVar; SCcaller; SCAN_SNV; SCIΦ
• VarTrix; SCmut
Genotype/eQTL (scRNA-seq data)
• SC-eQTL
Reconstruction of TCR/BCR sequences (scRNA-seq data)
• TraCer; BraCer; BALDR; VDJPuzzle; TRUST4
TCR/BCR integrative analysis (scTCR-seq/scBCR-seq data)
• scRepertoire; CellaRepertorium
Cell states and signalling
Trajectory analysis (scRNA-seq data)
• Monocle suite; PAGA; Slingshot; STEMNET; Scorpius
RNA velocity (scRNA-seq data)
• Velocyto; CellRank; scVelo
Pathway analysis (scRNA-seq data)
• PAGODA2; PLAGE; GSVA; Vision
Cell-to-cell communication (scRNA-seq data)
• CellPhoneDB; ICELLNET; NicheNet; CellChat
Gene regulatory network inference (scRNA-seq data)
• SCENIC; PIDC
Multi-modal and bulk tissue type composition
Joint multi-modal analysis (single-cell multi-omics data)
• WNN; TotalVI; CiteFuse; MOFA+
• LIGER; CCA
Bulk transcriptomics deconvolution (single-cell & bulk RNA seq data)
• MuSiC; RNASieve (inference of cellular composition)
• ScMappR; bMIND (cell type specific differential expression)
• CIBERSORTx; EcoTyper (transcriptome purification)
ST deconvolution (ST and scRNA-seq data)
• RCTD; SPOTlight; Stereoscope
• Cell2location; SpatialDWLS; CARD
ST neighbourhood analysis (ST data)
• Giotto; Squidpy; CellTrek
Target and biomarker discovery
Perturbation analysis (scCRISPR screening data)
• MIMOSCA; MUSIC;
• scMAGeCK; Mixscape
Cell-type prioritization (scRNA-seq/scATAC-seq data)
• Augur; MELD; Scissor


a Standard high-throughput chemical screens
• 100k–1M compounds
• Typically one compound dose tested on one condition
List of most active compounds
Dose–response and other studies
Hits suitable for drug discovery
b Single-cell high-throughput chemical screens
• 100–1,000 drugs
• Typically 5–10 drug doses tested on several cell lines (~100k–1M single cells in all)
Various perturbations
Pooling scRNA-seq
t-SNE2
t-SNE1
Treatment
Perturb A
Perturb B
Perturb C
Perturb D
Perturb A
Perturb B
Perturb C
Perturb D
Dose
Vehicle
Hints on drug's MoA


a
Patient enrolment
Sample collection
Tissue biopsy
Liquid biopsy
Single-cell sequencing
t-SNE2
t-SNE1
Association of cell type/ state abundance or genes with drug response
Cellular computational analysis
Responders
Proportion
1.00
0.75
0.50
0.25
0
Non-responders
Differential gene expression
Identification of cellular states or gene signature associated with the response to therapy
Patient stratification
Responders
Non-responders
b
Tumour biopsies from melanoma patients as samples
Responses to ICI treatments: both non-responding and responding tumours studied
Isolation of CD45+ immune cells and single-cell RNA sequencing
t-SNE2
t-SNE1
ICI response
ICI no-response
Differential expression analysis
CD8+ TCF7+ cells in ICI responders

HHS Public Access
Author manuscript
Trends Pharmacol Sci. Author manuscript; available in PMC 2021 December 01.

Published in final edited form as:
Trends Pharmacol Sci. 2020 December ; 41(12): 1050–1065. doi:10.1016/j.tips.2020.10.004.

Single-cell Techniques and Deep Learning in Predicting Drug Response

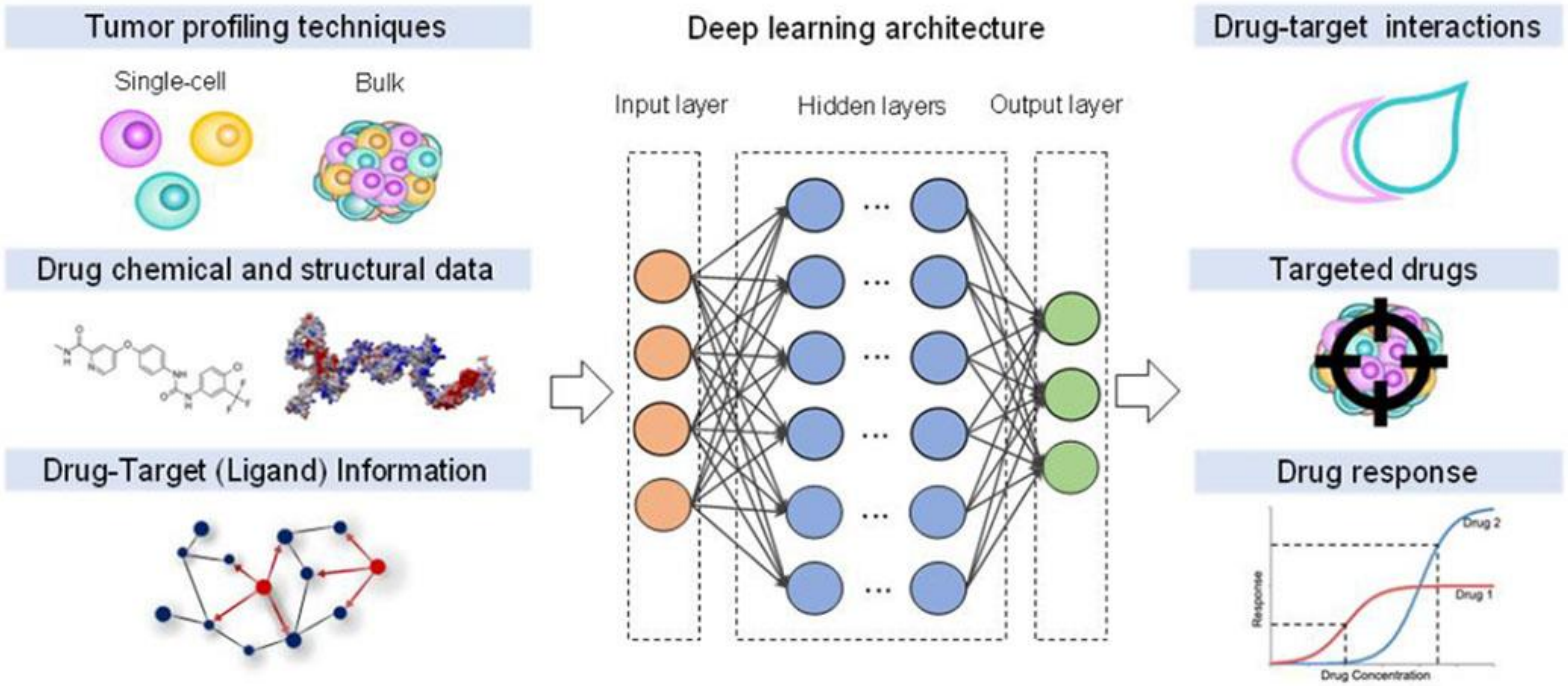
Zhenyu Wu[1,4], **Patrick J. Lawrence**[1,4], **Anjun Ma**[1], **Jian Zhu**[2], **Dong Xu**[3], **Qin Ma**[1,$]

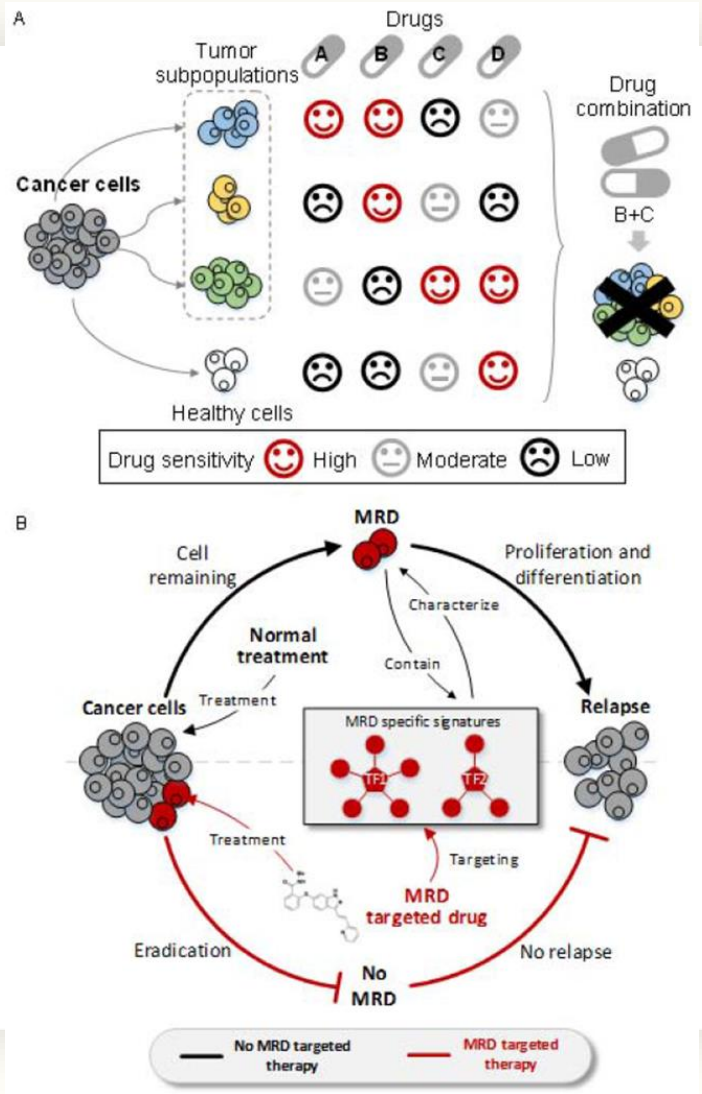
[1]Department of Biomedical Informatics, The Ohio State University, OH, 43235, USA

[2]Department of Pathology, The Ohio State University, OH, 43235, USA

[3]Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

[4]These authors contributed equally to the paper as first authors


Tumor profiling techniques
Single-cell
Bulk
Drug chemical and structural data
Drug-Target (Ligand) Information
Deep learning architecture
Input layer
Hidden layers
Output layer
Drug-target interactions
Targeted drugs
Drug response
Response
Drug 2
Drug 1
Drug Concentration


A
Drugs
Tumor subpopulations
A
B
C
D
Drug combination
B+C
Cancer cells
Healthy cells
Drug sensitivity
High
Moderate
Low
B
MRD
Cell remaining
Proliferation and differentiation
Characterize
Normal treatment
Contain
Treatment
Cancer cells
MRD specific signatures
TF1
TF2
Relapse
Treatment
Targeting
MRD targeted drug
Eradication
No relapse
No MRD
No MRD targeted therapy
MRD targeted therapy

nature communications

Article

https://doi.org/10.1038/s41467-022-34277-7

Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data

Received: 6 August 2021

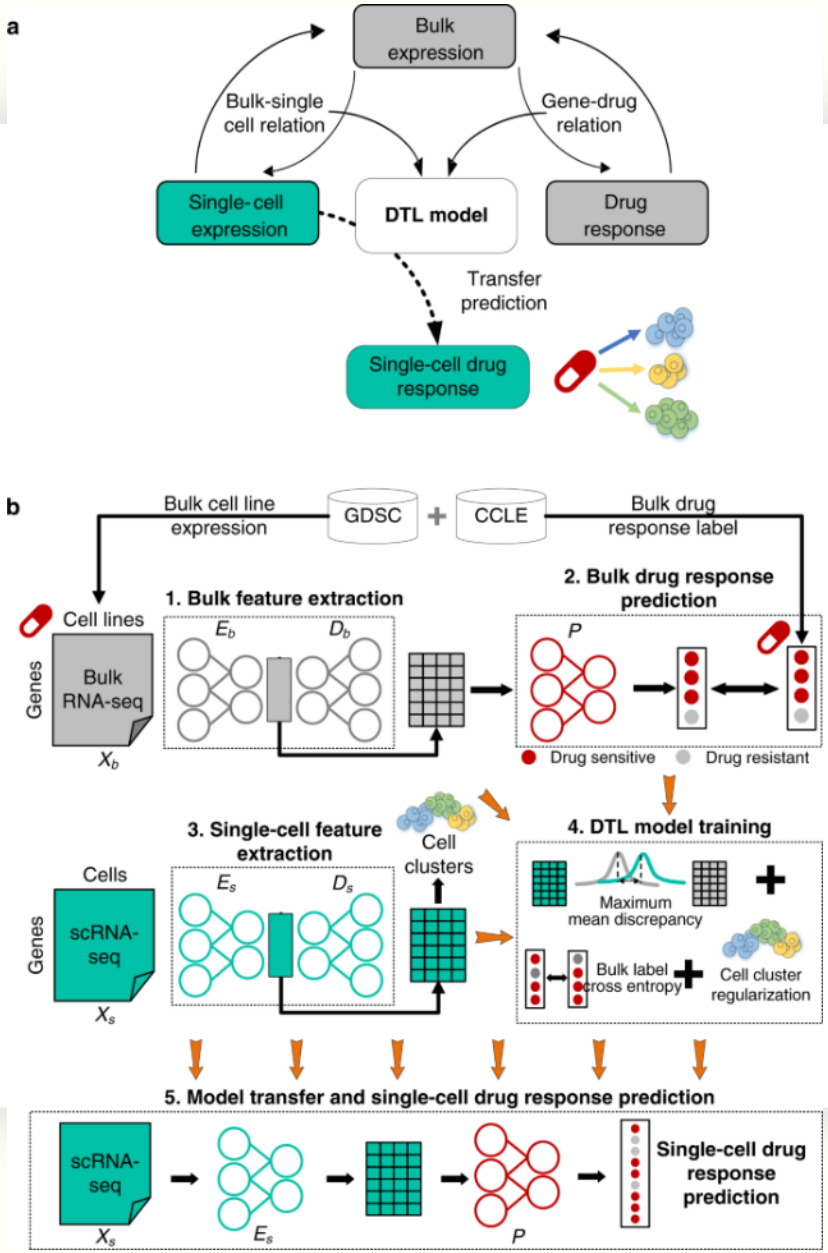
Accepted: 19 October 2022

Published online: 30 October 2022

Check for updates

Junyi Chen[1,6], Xiaoying Wang[2,6], Anjun Ma[1,3] ✉, Qi-En Wang[4], Bingqiang Liu[2], Lang Li[1], Dong Xu[5] & Qin Ma[1,3] ✉

Drug screening data from massive bulk gene expression databases can be analyzed to determine the optimal clinical application of cancer drugs. The growing amount of single-cell RNA sequencing (scRNA-seq) data also provides insights into improving therapeutic effectiveness by helping to study the heterogeneity of drug responses for cancer cell subpopulations. Developing computational approaches to predict and interpret cancer drug response in single-cell data collected from clinical samples can be very useful. We propose scDEAL, a deep transfer learning framework for cancer drug response prediction at the single-cell level by integrating large-scale bulk cell-line data. The highlight in scDEAL involves harmonizing drug-related bulk RNA-seq data with scRNA-seq data and transferring the model trained on bulk RNA-seq data to predict drug responses in scRNA-seq. Another feature of scDEAL is the integrated gradient feature interpretation to infer the signature genes of drug resistance mechanisms. We benchmark scDEAL on six scRNA-seq datasets and demonstrate its model interpretability via three case studies focusing on drug response label prediction, gene signature identification, and pseudotime analysis. We believe that scDEAL could help study cell reprogramming, drug selection, and repurposing for improving therapeutic efficacy.


a
Bulk expression
Bulk-single cell relation
Gene-drug relation
Single-cell expression
DTL model
Drug response
Transfer prediction
Single-cell drug response
b
Bulk cell line expression
GDSC
CCLE
Bulk drug response label
1. Bulk feature extraction
2. Bulk drug response prediction
Cell lines
Genes
Bulk RNA-seq
X_b
E_b
D_b
P
Drug sensitive
Drug resistant
3. Single-cell feature extraction
Cell clusters
4. DTL model training
Cells
scRNA-seq
X_s
E_s
D_s
Maximum mean discrepancy
Bulk label cross entropy
Cell cluster regularization
5. Model transfer and single-cell drug response prediction
Single-cell drug response prediction

Research
A SCIENCE PARTNER JOURNAL

REVIEW

Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level

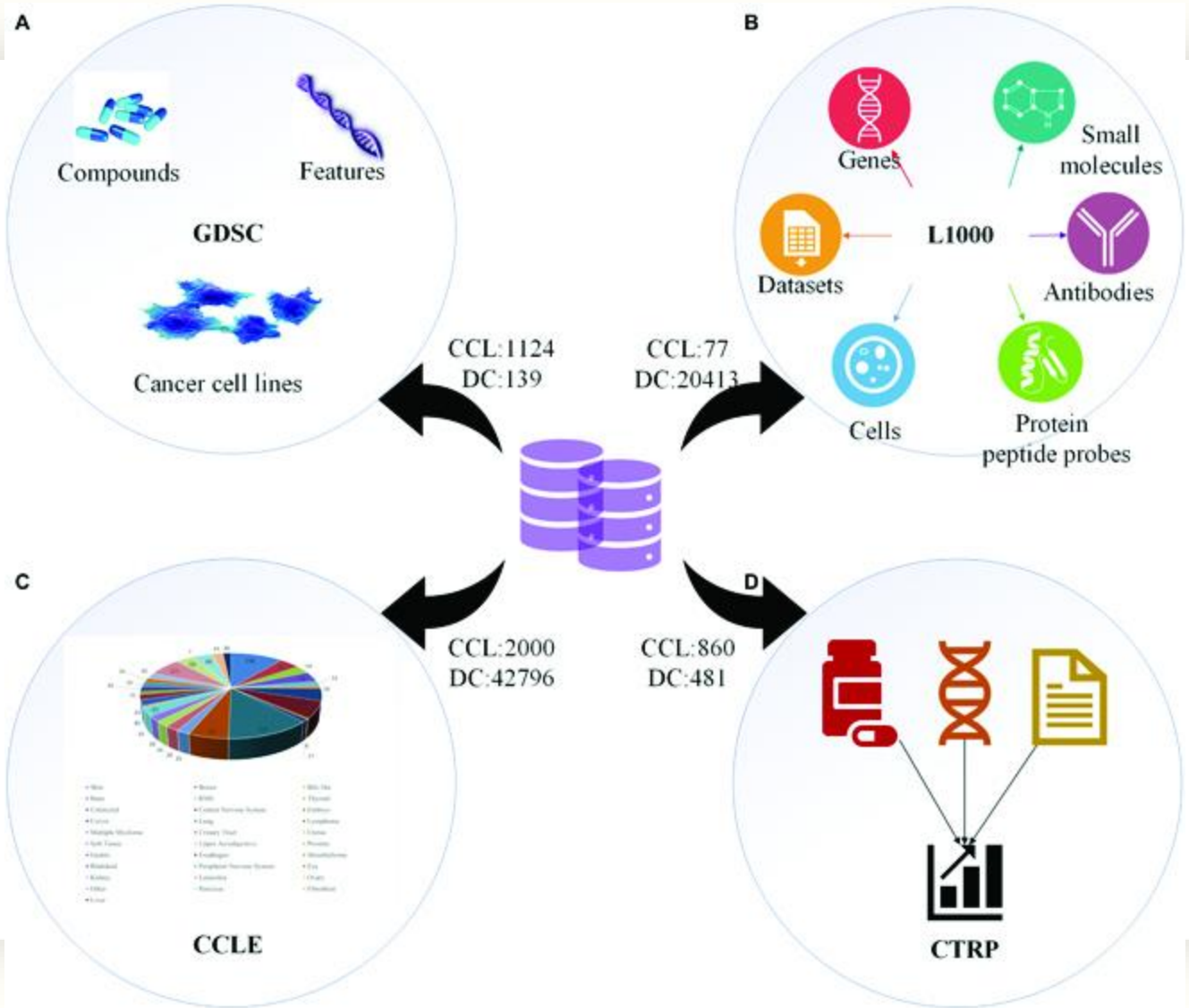
Ren Qi[1,2] and Quan Zou[1,3*]

[1]Yangtze Delta Region Institute (Quzhou), University of Electronic Science and Technology of China, Quzhou 324000, China. [2]School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China. [3]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China.

*Address correspondence to: zouquan@nclab.net

Citation: Qi R, Zou Q. Trends And Potential Of Machine Learning And Deep Learning In Drug Study At Single-Cell Level. *Research* 2023;6:Article 0050. https://doi.org/10.34133/research.0050

Submitted 22 September 2022
Accepted 27 December 2022
Published 9 March 2023


A
Compounds
Features
GDSC
Cancer cell lines
B
Genes
Small molecules
L1000
Datasets
Antibodies
Cells
Protein peptide probes
CCL:1124
DC:139
CCL:77
DC:20413
C
CCL:2000
DC:42796
CCLE
D
CCL:860
DC:481
CTRP

Inputs
DTI
Protein structure
Protein sequence
PPI
Drug similarity
Target similarity
SMILE
Genomic sequence
GB
GA
AMP
DDI
DCI
Image data
Copy number
DNA methylation
Models
CNN
RNN
GCN
DTL
DNN
VAE
Bayesian
Gaussian likelihood
HNB
Kernel methods
Ensemble learning
Transform
Results
Drug response
Drug discovery
DTI
Other
DeepConv-DTI
BMTMKL
HNMDRP
Dr.VAE
DeepDTnet
deepDR
DeepDTIs
DeepCPI
DeepDRK
AtomNet
MOLI
Decagon
DEEPScreen
DeepAffinity
DeepPurpose
DeepChem
DeepSynergy
tCNNs
GraphDRP
RefDNN
DeepCDR
DCell
Deep-AmPEP30
DeepDR
DeepMACT
Methods

Seminars in Cancer Biology 91 (2023) 50–69

Contents lists available at ScienceDirect

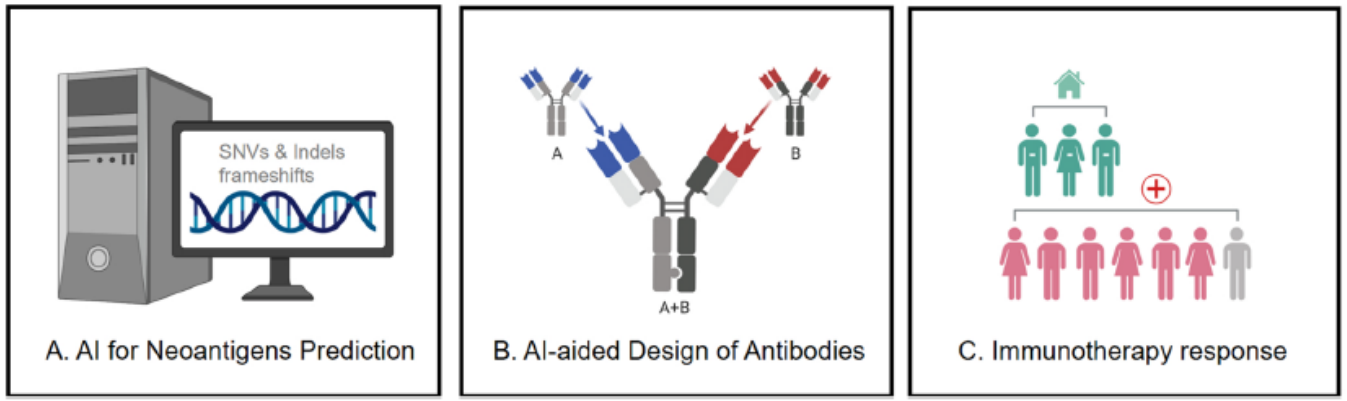
Seminars in Cancer Biology

journal homepage: www.elsevier.com/locate/semcancer

Artificial intelligence in cancer immunotherapy: Applications in neoantigen recognition, antibody design and immunotherapy response prediction

Tong Li [a,1], Yupeng Li [a,1], Xiaoyi Zhu [b,d,1], Yao He [a], Yanling Wu [b,d], Tianlei Ying [b,d,*,1], Zhi Xie [a,c,**]

[a] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China
[b] MOE/NHC Key Laboratory of Medical Molecular Virology, Shanghai Institute of Infectious Disease and Biosecurity, School of Basic Medical Sciences, Shanghai Medical College, Fudan University, Shanghai, China
[c] Center for Precision Medicine, Sun Yat-sen University, Guangzhou, China
[d] Shanghai Engineering Research Center for Synthetic Immunology, Shanghai, China


SNVs & Indels
frameshifts
A. AI for Neoantigens Prediction
A
B
A+B
B. AI-aided Design of Antibodies
C. Immunotherapy response

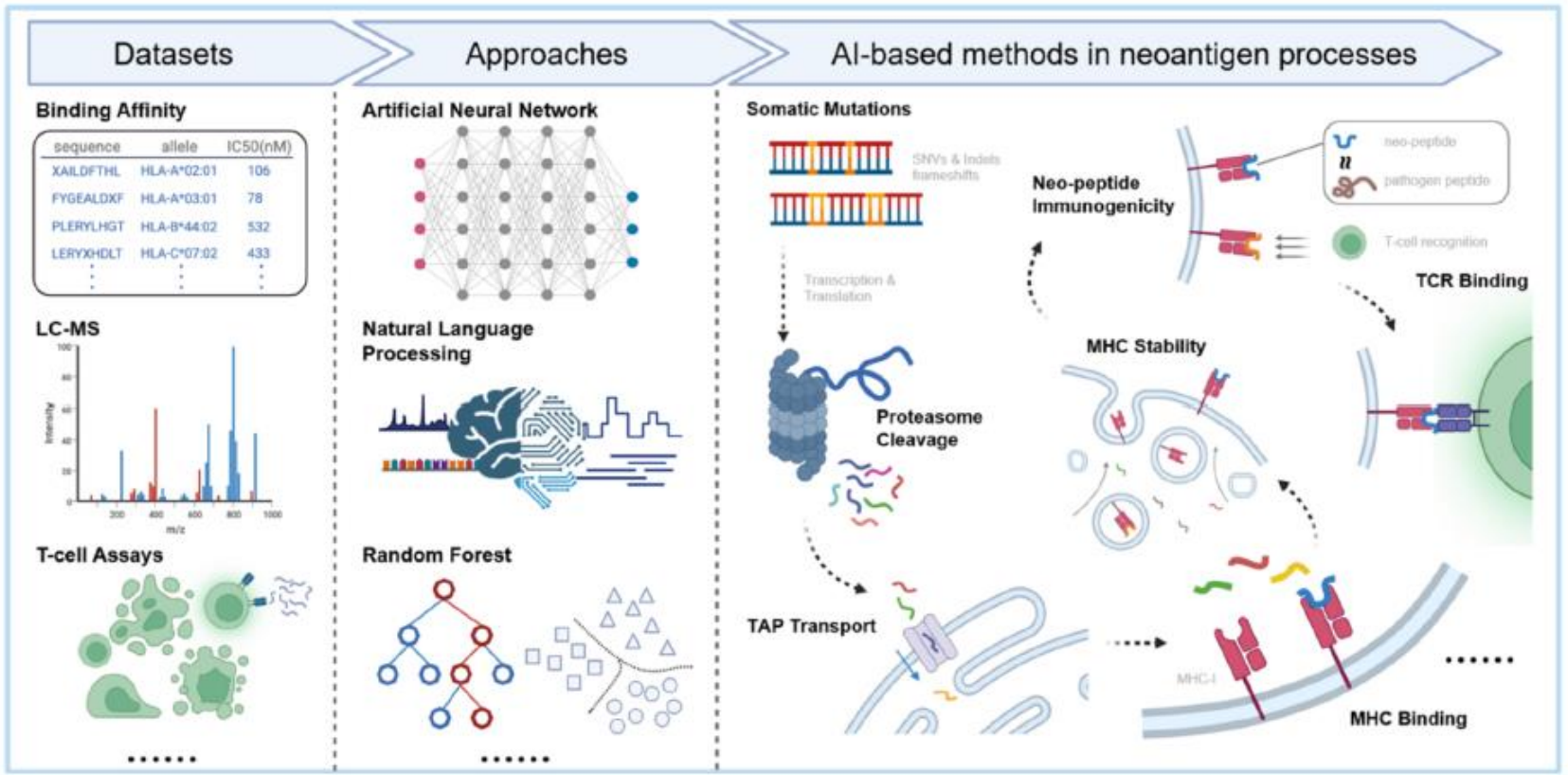
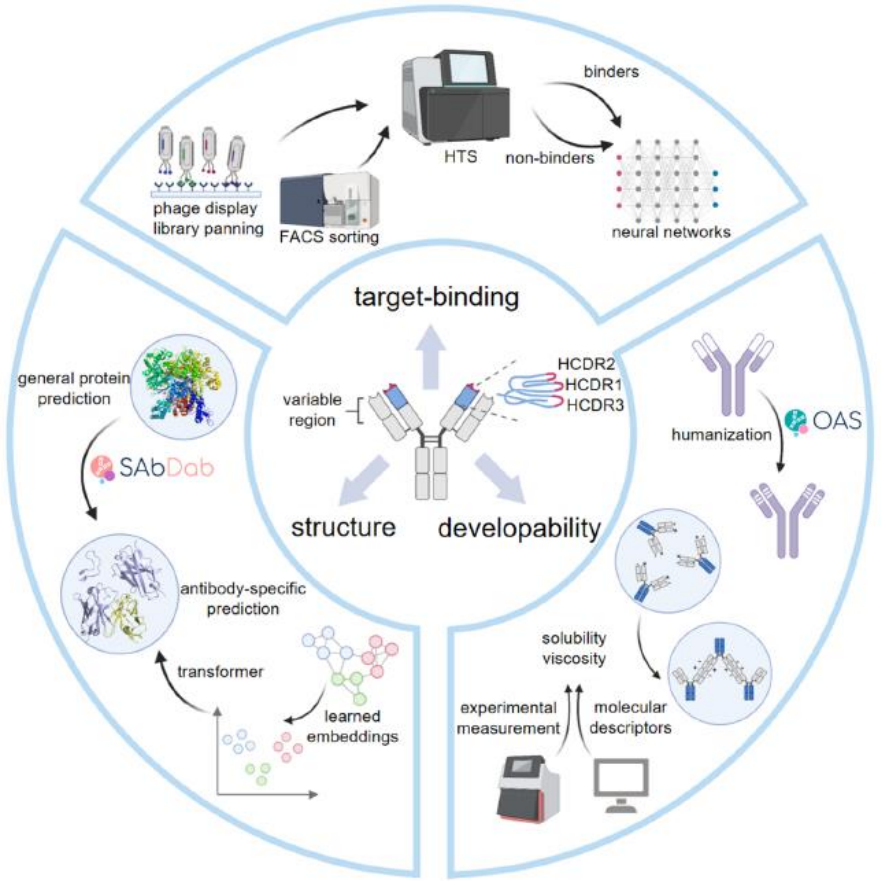

Datasets
Approaches
AI-based methods in neoantigen processes
Binding Affinity
sequence allele IC50(nM)
XAILDFTHL HLA-A*02:01 106
FYGEALDXF HLA-A*03:01 78
PLERYLHGT HLA-B*44:02 532
LERYXHDLT HLA-C*07:02 433
LC-MS
T-cell Assays
Artificial Neural Network
Natural Language Processing
Random Forest
Somatic Mutations
SNVs & Indels frameshifts
Transcription & Translation
Proteasome Cleavage
TAP Transport
Neo-peptide Immunogenicity
neo-peptide
pathogen peptide
T-cell recognition
TCR Binding
MHC Stability
MHC-I
MHC Binding

Table 1
Application of different AI algorithms in immunotherapy.

1. AI-based methods for neoantigen prediction

Name	Year	MHC	Approach	Neoantigen process	Type of learning data	Training sets	Independent benchmark
EDGE [38]	2019	I/II	DL	MHC binding	MHC ligands	142,844 peptides from 101 samples	
DeepHLApan [225]	2019	I	RNN	MHC binding & TCR binding	MHC ligands & T-cell recognized neoepitopes	437,077 peptides for binding model & 32,785 peptides for immunogenic	
NMER [63]	2021	I	RF	TCR binding	T-cell recognized neoepitopes	185 neoepitopes identified by TILs from 96 samples	
NetMHC-4.0 [47]	2016	I	NN	MHC binding	Affinity measurements	Affinity measurements from IEDB	[44,46]
NetMHCpan-4.0 [36]	2017	I	NN	MHC binding	Affinity measurements & MHC ligands	Affinity measurements from IEDB & 85,217 peptides	[44,46]
NetMHCstabpan [67]	2016	I	NN	pMHC stability	Half-life of pMHC in vitro	28,166 binding measurements	[44]
MHCflurry [50]	2018	I	NN	MHC binding	Affinity measurements & MHC ligands	230,735 affinity measurements & 226,684 peptides	[44,46]
MHCflurry-2.0 [51]	2020	I	NN	binding affinity & MHC ligands	Affinity measurements & MHC ligands	219,596 affinity measurements & 493,473 peptides for binding model; 399,392 peptides for presentation model	
Neonmhc2 [61]	2019	II	CNN	MHC binding	MHC ligands	> 50,000 peptides	
Neopesee [226]	2018	I	ML	Immunogenicity	Epitopes with T-cell response	311 neoepitopes with T-cell response	
pMTnet [227]	2021	I	DL	Immunogenicity	Affinity measurements & pMHC-TCR pairs	172,422 affinity measurements & 243,747 human TCRβ CDR3 sequences & 32,607 pMHC-TCR pairs	
ForestMHC [228]	2019	I	RF	MHC binding	9-mer peptides from MS	> 160,000 peptides	
PRIME [96]	2021	I	LR	MHC binding & TCR binding	Peptides with immunogenicity or not	4958 peptides	[229]
MARIA [54]	2019	II	RNN	MHC binding & protease cleavage & expression	MHC ligands	8374 peptides for presentation model & 33,909 peptides for binding model & 12,150 for cleavage model	
MHCSeqNet [60]	2019	I	NLP, GRU	MHC binding & ligand prediction	MHC ligands	228,348 peptides	
HLAthena [230]	2019	I	NN	MHC binding	MHC ligands	> 185,000 peptides	
NetTCR-2.0 [231]	2021	I/II	CNN	Immunogenicity	TCR-peptide pairs	9204 CDR3β-peptide pairs & 4598 CDR3α-/β-peptide pairs	
NetMHCPan-4.1 [232]	2020	I	NN	MHC binding	Affinity measurements & MHC ligands	208,093 affinity measurements & 665,492 peptides	
NetMHCIIpan-4.0 [232]	2020	II	NN	MHC binding	Affinity measurements & MHC ligands	108,959 affinity measurements & 381,066 peptides	

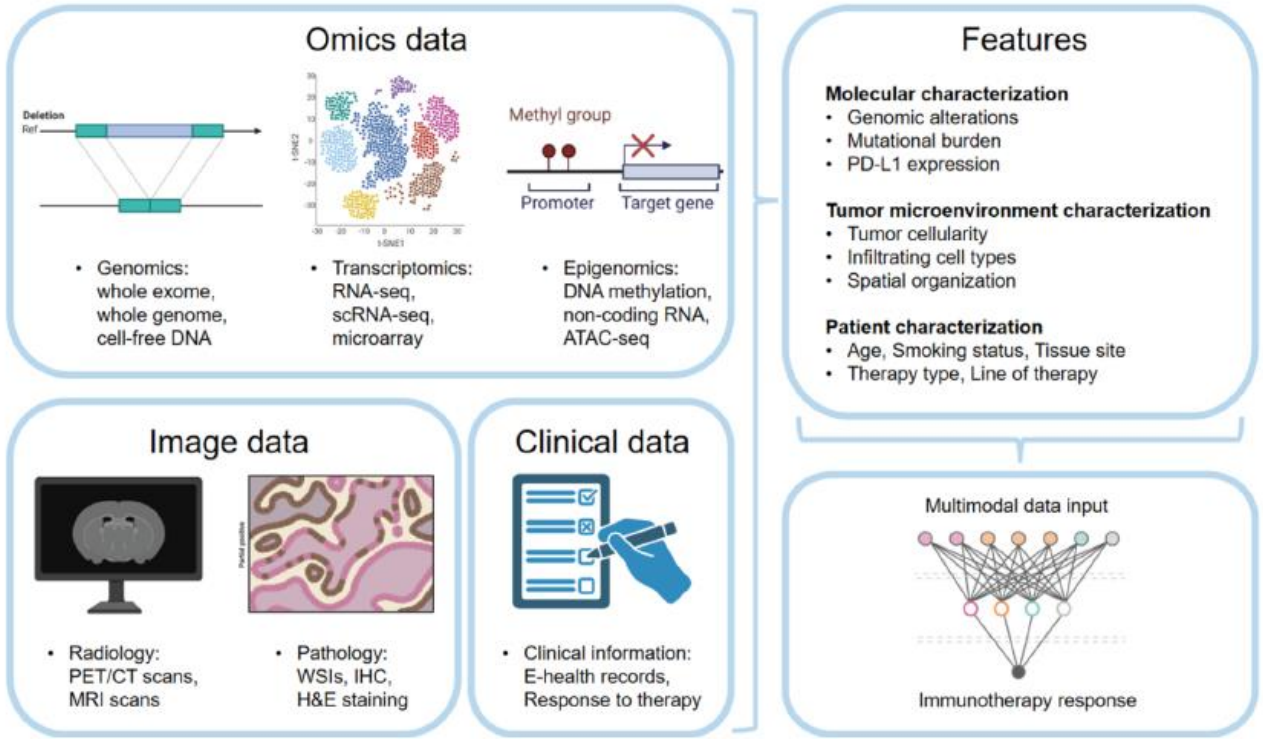

phage display
library panning
FACS sorting
HTS
binders
non-binders
neural networks
target-binding
variable
region
HCDR2
HCDR1
HCDR3
structure
developability
general protein
prediction
SAbDab
antibody-specific
prediction
transformer
learned
embeddings
humanization
OAS
solubility
viscosity
experimental
measurement
molecular
descriptors

2. AI-based methods for antibody design

Role	Name	Model	Training input	Description	Ref.
Target binding prediction and optimization	Ens-Grad	Ensemble of CNN	51,130 CDRH3 sequences and their R2-to-R3 enrichment in phage-display panning	Designing antibody CDRH3 regions with target affinities and improved specificity	[107]
	-	CNN	CDRH3 sequences of 11,300 binding and 27,539 non-binding trastuzumab variants	Optimization of trastuzumab in affinity and pharmaceutical properties	[108]
	-	CNN and generative adversarial network	6003 non-binder and 1345 binder sequences for CTLA-4; 6052 non-binder and 1719 binder sequences for PD-1	Classifying and generating sequences of CTLA-4 and PD-1 binding antibodies	[109]
	-	LSTM	959 VH sequences	Generating antibody sequences with improved affinity to kynurenine	[110]
	-	linear discriminant analysis model	4000 sequences (2000/2000 with high and low specificity; 1516/2484 with high and low antigen binding)	Co-optimizing the affinity and specificity of emibetuzumab	[111]
Antibody structure prediction	DeepH3	Series of CNN	1388 structures from SAbDab	Predicting inter-residue distances and orientations of antibody CDRH3 region	[116]
	DeepAb	ReNet and bi-LSTM encoder	118,386 paired sequences from the OAS database and 1692 structures from SAbDab	Predicting relative distances and orientations of antibody Fv region and realizing structure by Rosetta	[121]
	DeepSCAb	ResNet for the inter-residue module and a transformer encoder model for the rotamer module	1433 structures from SAbDab	Predicting full Fv structures including side-chain geometries	[126]
	ABlooper	five E(n)-equivariant graph neural networks	3438 structures from SAbDab	Predicting structures of antibody CDR loops	[127]
	NanoNet	Two 2D CNN	~ 2000 heavy chains of mAbs and nanobody structures	Modeling nanobodies and VH domains of antibodies	[130]
Pharmaceutical property	AbLSTM	Bi-LSTM network	25,000 antibody sequences from BCR repertoire sequencing	Evaluating the nativeness of antibody candidates	[135]
	BioPhi	RoBERTa transformer encoder model for Sapiens	Human antibody repertoires in the OAS database	A platform composed of OASis which is for antibody humanness evaluation and Sapiens for antibody humanization	[136]
	solPredict		Protein solubility data of 220 antibodies		[138]

(*continued on next page*)

Role	Name	Model	Training input	Description	Ref.
		Support vector machine and random forest models with language-based transfer learning		Predicting the apparent solubility of antibodies in histidine (pH 6.0) buffer	
	DeepSCM	1D CNN	6596 antibody Fv sequences and their SCM scores	A surrogate model for the SCM to predict antibody viscosity	[140]
	-	Random forest classifier	64 clinical-stage antibodies and their PK data	Identifying biophysical assays and in silico properties correlated with antibody PK	[144]


Omics data
Deletion
Ref
Methyl group
Promoter
Target gene
• Genomics: whole exome, whole genome, cell-free DNA
• Transcriptomics: RNA-seq, scRNA-seq, microarray
• Epigenomics: DNA methylation, non-coding RNA, ATAC-seq
Image data
• Radiology: PET/CT scans, MRI scans
• Pathology: WSIs, IHC, H&E staining
Clinical data
• Clinical information: E-health records, Response to therapy
Features
Molecular characterization
• Genomic alterations
• Mutational burden
• PD-L1 expression
Tumor microenvironment characterization
• Tumor cellularity
• Infiltrating cell types
• Spatial organization
Patient characterization
• Age, Smoking status, Tissue site
• Therapy type, Line of therapy
Multimodal data input
Immunotherapy response

3. AI-based methods for immunotherapy response prediction

	Target	Summary	Data type	Tumor type	Technology	Validation	Ref.
Prediction of associated biomarker	MSI	Predict MSI directly from H&E histology	H&E histology	Gastrointestinal cancer	CNN with deep residual learning (resnet18)	AUC = 0.84	[158]
	MSI or dMMR	Identified dMMR or MSI from H&E-stained slides	H&E-stained slides	Colorectal cancer	Modified ShuffleNet/ MobileNetV2 architecture	AUPRC = 0.92–0.931, 66.6–67 % SPE and 76.0–95 % SEN	[161, 163]
	PD-L1 score	Quantitatively analyze the PD-L1 score of tumor cells	PD-L1-stained digital images	Cutaneous melanoma	ML algorithm (random forest classifier)	Correlation between label and prediction ($r = 0.97$, $P < 0.0001$).	[164]
	PD-L1 expression	Predict the TPS of PD-L1 expression	WSIs of the 22c3 assay	Non-small cell lung cancer	U-Net structure with residual blocks	ACC = 0.9326 and SPE = 0.9641	[166]
	TMB	Predicting TMB from histopathological images	H&E-stained histopathological slides	Lung adenocarcinoma	Inception-v3 and the random forest architecture	AUPRC = 0.92, precision = 0.89	[168]
	TMB status	Predicting TMB from images and clinical information	Histopathological images and clinical data	Colorectal cancer	Multi-modal deep learning model based on ResNet	AUPRC = 0.817	[172]
Deciphering TME	TME reconstruction	Prediction of 51 unique cell subpopulations	Bulk RNA-seq	Pan-cancer from TCGA	Decision tree ML deconvolution algorithm	Cytometric, immunohistochemical, or scRNA-seq	[182]
	TILs maps	Predict spatial patterns of TILs	H&E images	13 TCGA tumor types	Semi-supervised CNN and unsupervised CAE	Prediction correlate with pathologist and molecular estimates	[189]
	Tumor Cellularity	Predict tumor purity from H&E Image	Whole Slide H&E Image	Breast cancer	Weakly-supervised segmentation model with Resnet-34	Cohen's kappa coefficient = 0.69	[193]
	TME	Predict spatial mapping of multiple cell types	H&E images and spatial transcriptomic data	Lung adenocarcinoma	CNN	Correlated with bulk RNA-seq	[202]
Directly prediction	Immunotherapy response	Predict ICI treatment responses in three different cancer types	Clinical outcome and transcriptomic data	Melanoma, gastric cancer, and bladder cancer	Protein—protein interaction network (PPI)-based ML	AUCs > 0.7	[207]
	Immunotherapy response	A predictive model of immunotherapy using genomics data	Genomic data and RNA-seq	29 cancer types from TCGA	MultiModal Network	Compared to benchmark markers (p = 0.009)	[208]
	ICB efficacy	Predict ICB response	Genomic, molecular, demographic and clinical data	16 different cancer types	Ensemble learning random forest ML model	Pan-cancer AUC = 0.85	[209]
	Immunotherapy response	Prediction of response to PD-(L)1 blockade	Medical imaging, histopathological, and genomic features	Non-small cell lung cancer	Multiple-instance LR and multimodal dynamic attention	AUC = 0.80	[216]

Table 2
Datasets for different applications of immunotherapy.

1. Datasets for neoantigen prediction

Name	Year	Last updated	Overview of the data
IEDB [99]	2004	2023/1/30	> 1,500,000 epitopes from > 23,000 studies, including B cell, T cell, and MHC ligand assays
dbPepNeo2 [233]	2022	No record	801 HC neoantigens and 864,884 low LC peptidomes, validated neoantigen peptides, TCRs, and HLA peptidomes

2. Datasets for antibody design

Database name	Description	Link	Ref.
Observed Antibody Space (OAS)	Collecting immune repertoires	https://opig.stats.ox.ac.uk/webapps/oas/	[122]
Structural Antibody Database (SAbDab)	Database of antibody structures	https://opig.stats.ox.ac.uk/webapps/newsabdab/sabdab/	[117]

3. Datasets for immunotherapy response prediction

Dataset name	Description	Link	Type(s) of cancer (s)	Ref.
Predict MSI in gastrointestinal cancer	Histological images for MSI vs. MSS classification in gastrointestinal cancer	https://doi.org/10.5281/zenodo.2530789 https://doi.org/10.5281/zenodo.2530835 https://doi.org/10.5281/zenodo.2532612	Gastrointestinal cancer	[158]
The Cancer Genome Atlas (TCGA)	TCGA is a pool of molecular data sets publicly accessible and freely available to cancer researchers	https://portal.gdc.cancer.gov/	Pan-cancer	[161]
NCT-CRC-HE-100K and CRC-VAL-HE-7K datasets	This is a set of 100,000 image patches from H&E stained histological images of colorectal cancer (CRC) and normal tissue.	https://doi.org/10.5281/zenodo.1214455	Colorectal cancer	[163]
MMDL-colorectal cancer	Predicting colorectal cancer TMB from histopathological images and clinical information	https://github.com/hkmgeneis/MMDL/tree/master	Colorectal cancer	[172]
Kassandra	Precise reconstruction of the TME using bulk RNA-seq and a ML algorithm trained on artificial transcriptomes	https://science.bostongene.com/kassandra/	Pan-cancer	[182]
CIBERSORTx	Determining cell type abundance and expression from bulk tissues with digital cytometry	http://cibersortx.stanford.edu/	Pan-cancer	[183]
Scaden	DL–based cell composition analysis from tissue expression profiles	https://www.science.org/doi/10.1126/sciadv.aba2619#sec-4	Pan-cancer	[184]
MethylCIBERSORT	Pan-cancer deconvolution of tumor composition using DNA methylation	https://www.nature.com/articles/s41467-018-05570-1#Sec12	Pan-cancer	[185]
MethylNet	An automated and modular deep learning approach for DNA methylation analysis	https://doi.org/10.24433/CO.6373790.v1	Pan-cancer	[186]
Human-interpretable image features (HIFs)	HIFs derived from densely mapped cancer pathology slides predict diverse molecular phenotypes	https://www.nature.com/articles/s41467-021-21896-9#data-availability	Pan-cancer	[194]
NetBio	Network-based ML approach to predict immunotherapy response in cancer patients	https://zenodo.org/record/4661265 http://research-pub.gene.com/IMvigor210CoreBiologies/https://string-db.org/	Pan-cancer	[207]
DeepTCR	DL reveals predictive sequence concepts within immune repertoires to immunotherapy	https://zenodo.org/record/6590069	Pan-cancer	[210]
PD-(L)1 blockade in patients with NSCLC	Multimodal integration of radiology, pathology and genomics for prediction of response	https://www.synapse.org/#!Synapse:syn26642505 https://www.cbioportal.org/study/summary?id=lung_msk_mind_2020	Non-small cell lung cancer	[216]

Received: 18 October 2019 | Revised: 26 November 2019 | Accepted: 16 December 2019

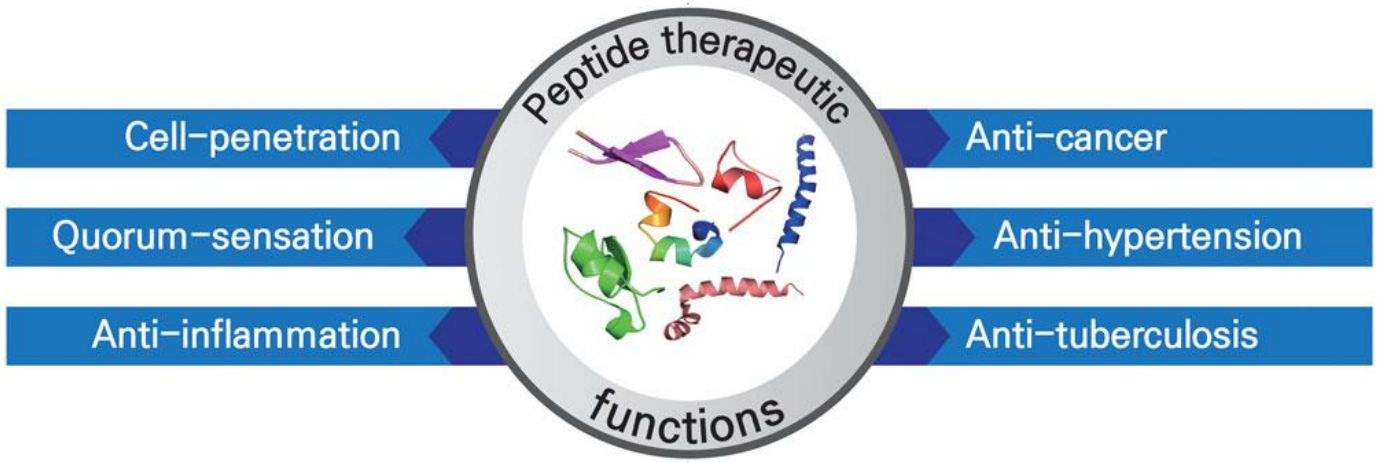
DOI: 10.1002/med.21658

REVIEW ARTICLE

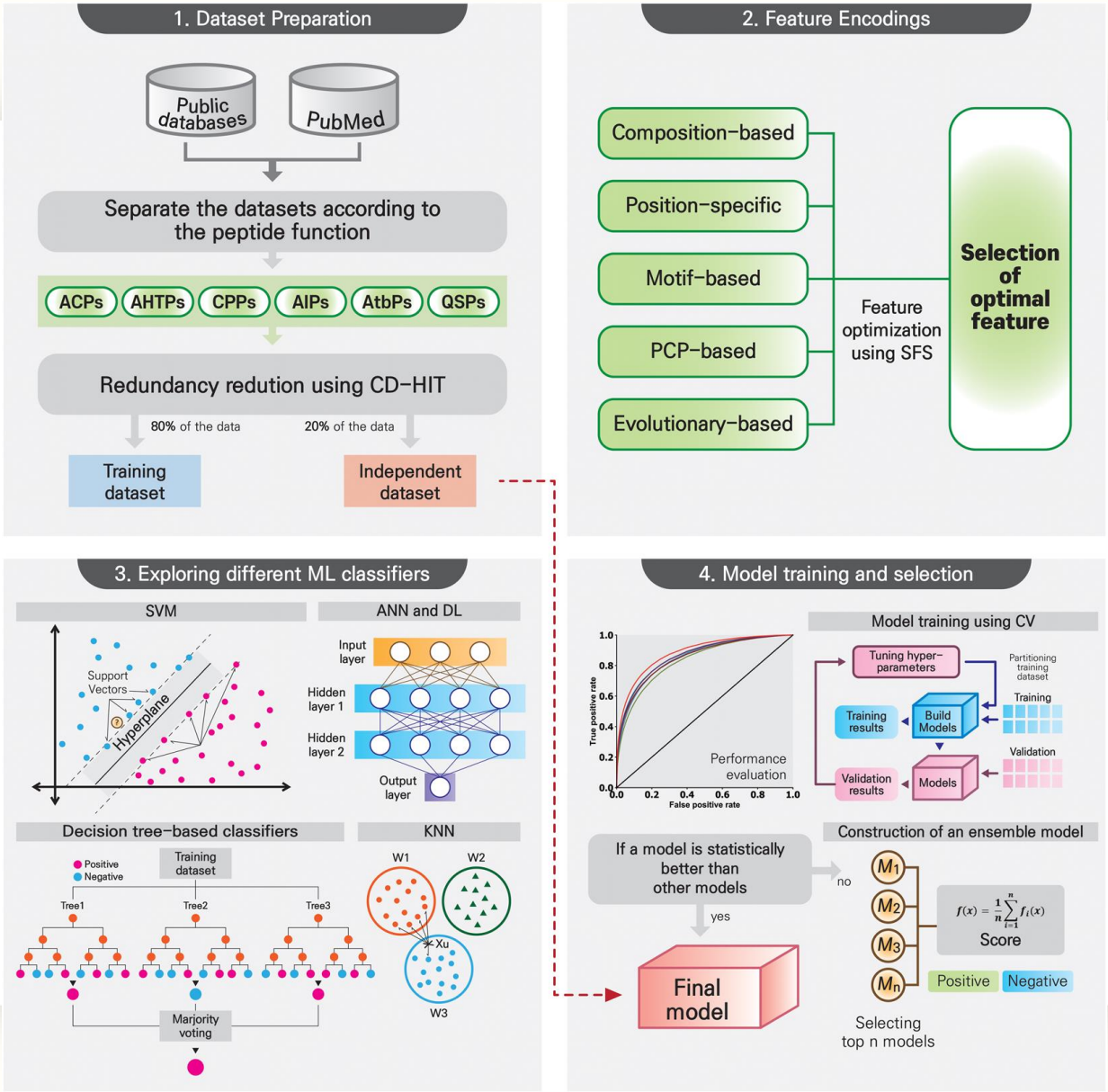
WILEY

Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening

Shaherin Basith | Balachandran Manavalan | Tae Hwan Shin | Gwang Lee



* Source: xxxx, 2022.0x.xx



* Source: xxxx, 2022.0x.xx

TABLE 2 List of currently available ACP predictors evaluated in this review

Predictor	Classifier, Year	Data set size (number of positive/ negative samples)	Feature encodings	Evaluation strategy	Accuracy (CV/ independent)	Web server availability
AntiCP[95]	SVM, 2013	225/225 *50/50*	BPF	10-Fold CV	0.914/0.89	http://crdd.osdd.net/raghava/anticp/
Hajisharifi et al[96]	SVM, 2013	138/206 —	Local alignment kernel	5-Fold CV	0.897/—	NA
ACPP[97]	SVM, 2015	217/3979 *40/40*	Protein-relatedness measure	10-Fold CV	0.97/0.962	†http://acpp.bicpu.edu.in/predict.php
iACP[98]	SVM, 2016	138/206 *150/150*	GGDPC	5-Fold CV	0.947/0.927	http://lin-group.cn/server/iACP
Li and Wang[99]	SVM, 2016	138/206 *150/150*	AAC, average chemical shifts, and RAAC	LOOCV	0.936/0.893	NA
Khan et al[100]	SVM, 2017	138/206 —	SAAC	LOOCV	0.933/—	NA
iACP-GAEnsC[101]	(SVM, RF, PNN, KNN, and GRNN), 2017	138/206 —	Amphiphilic PseAAC, GGDPC, and RAAC	LOOCV	0.965/—	NA
MLACP[12]	RF, 2017	187/398 *422/422*	AAC, ATC, DPC, and PCP	10-Fold CV	0.872/0.827	www.thegleelab.org/MLACP.html
SAP[102]	SVM, 2018	138/206 —	GGDPC	5-Fold CV	0.919/—	NA
ACPred-FL[67]	SVM, 2018	250/250 *82/82*	5 Class features	10-Fold CV	0.914/0.884	http://server.malab.cn/ACPred-FL
mACPpred[28]	SVM, 2019	266/266 *157/157*	AAC, DPC, CTD, QSO, AAI, BPF, and CTD	10-Fold CV	0.917/0.914	http://thegleelab.org/mACPpred
ACPred[103]	SVM, 2019	138/205 —	AAC and Am-PseAAC	LOOCV	0.956/—	†https://codes.bio/acpred/
ACP-DL[104]	DL (LSTM), 2019	376/364 —	BPF, *k*-mer sparse matrix	5-Fold CV	0.815/—	§https://github.com/haichengyi/ACP-DL

TABLE 7 List of currently available CPP predictors evaluated in this review

Predictor	Classifier, Year	Data set size (number of positive/ negative samples)	Feature encodings	Evaluation strategy	Accuracy (CV/ independent)	Web server availability
Dobchev et al[61]	ANN, 2010	49/10 *23/2*	QSAR descriptors	3-Fold CV	0.983/0.96	NA
Sanders et al[62]	SVM, 2011	111/34 —	PCP	10-Fold CV	0.917/—	NA
CellPPD[63]	SVM, 2013	708/708 *99/99*	A combination of motif and BPF	5-Fold CV	0.974/0.813	http://crdd.osdd.net/raghava/cellppd/
CPPpred[66]	N-to-1 NN, 2013	74/100 *47/47*	Motif	5-Fold CV	0.776/0.830	†http://bioware.ucd.ie/cpppred
C2Pred[115]	SVM, 2016	411/411 —	DPC	5-Fold CV	0.836/—	http://lin-group.cn/server/C2Pred
SkipCPP-Pred[67]	RF, 2017	462/462 —	Adaptive k-skip-2-gram	LOOCV	0.906/—	http://server.malab.cn/SkipCPP-Pred/Index.html
CPPred-RF[91]	RF, 2017	462/462 —	PseAAC, DPC and PCP	LOOCV	0.916/—	http://server.malab.cn/CPPred-RF
MLCPP[93]	ERT, 2018	427/427 311/311	AAC and PCP	10-Fold CV	0.883/0.896	www.thegleelab.org/MLCPP
KELM-CPPpred[116]	KELM, 2018	408/408 96/96	AAC and the motif-based features	10-Fold CV	0.870/0.870	http://sairam.people.iitgn.ac.in/KELM-CPPpred.html
CPPred-FL[92]	RF, 2018	462/462 —	19 Probabilistic features	10-Fold CV	0.921/—	http://server.malab.cn/CPPred-FL
G-DipC[117]	XGB, 2019	1223/1223 —	General DPC	5-Fold CV	0.985/—	NA
PEPred-SUITE[16]	RF, 2019	370/370 92/92	15 Probabilistic features	10-Fold CV	0.912/0.880	http://server.malab.cn/PEPred-Suite

JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/jacsau

Article

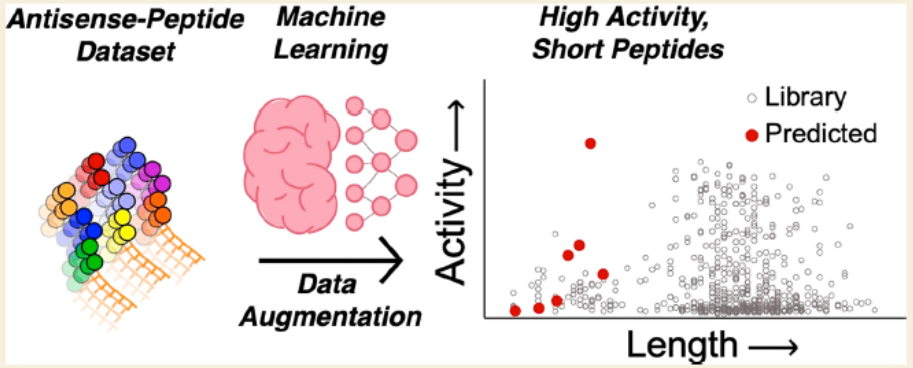
Deep Learning Enables Discovery of a Short Nuclear Targeting Peptide for Efficient Delivery of Antisense Oligomers

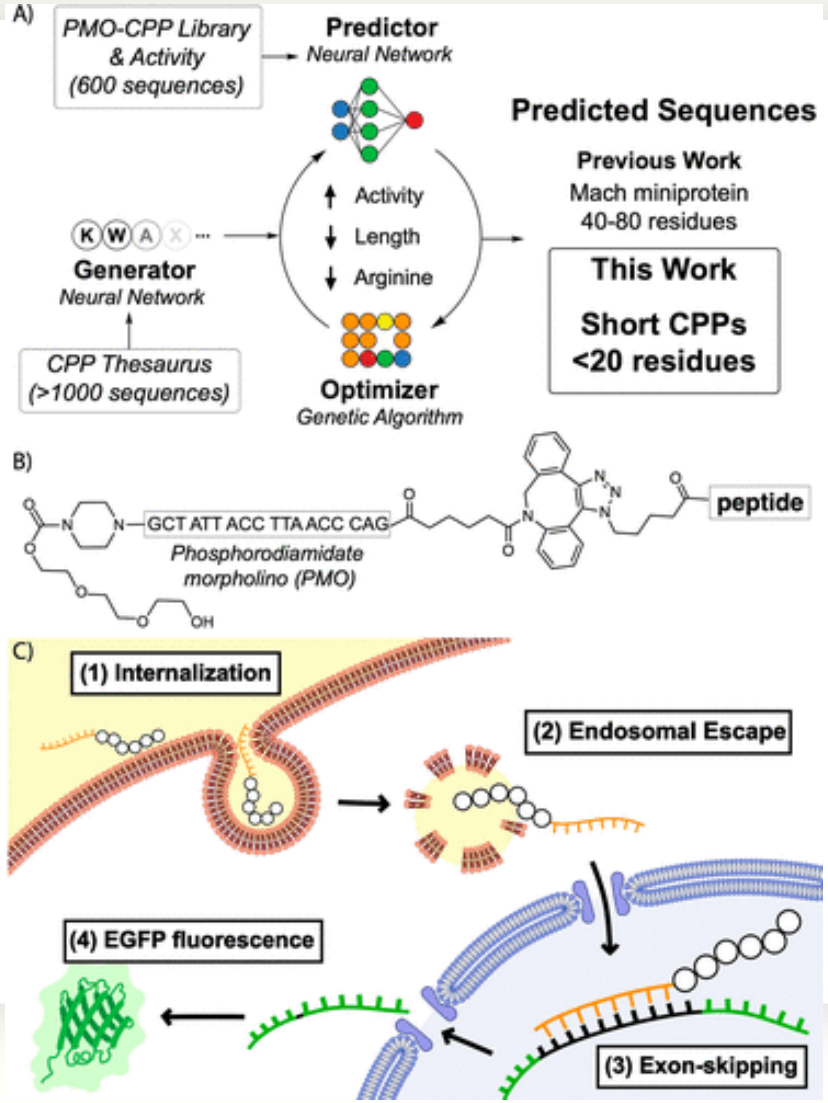
Eva M. López-Vidal,[▽] Carly K. Schissel,[▽] Somesh Mohapatra, Kamela Bellovoda, Chia-Ling Wu, Jenna A. Wood, Annika B. Malmberg, Andrei Loas, Rafael Gómez-Bombarelli,* and Bradley L. Pentelute*

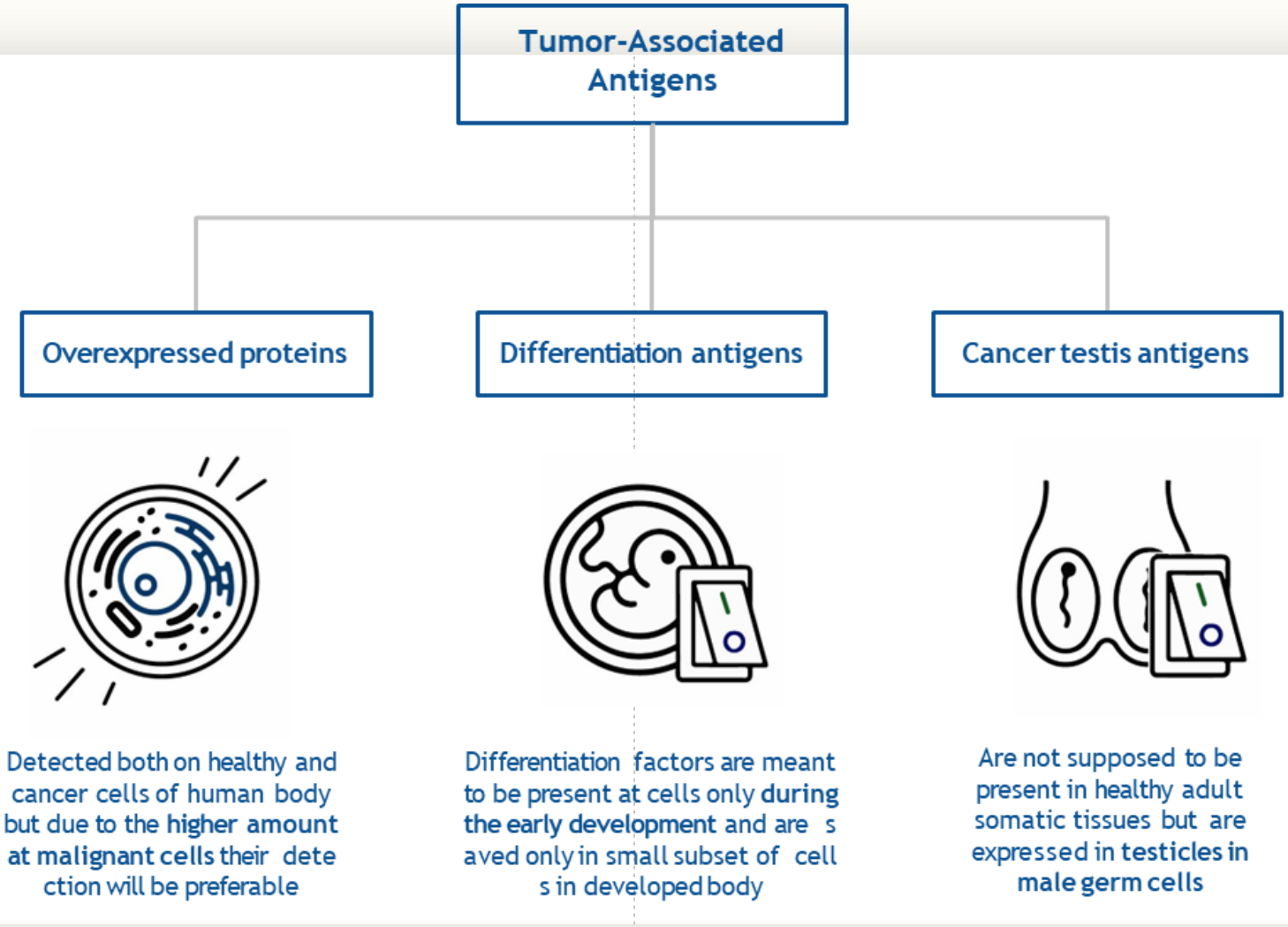
Cite This: *JACS Au* 2021, 1, 2009–2020

Read Online

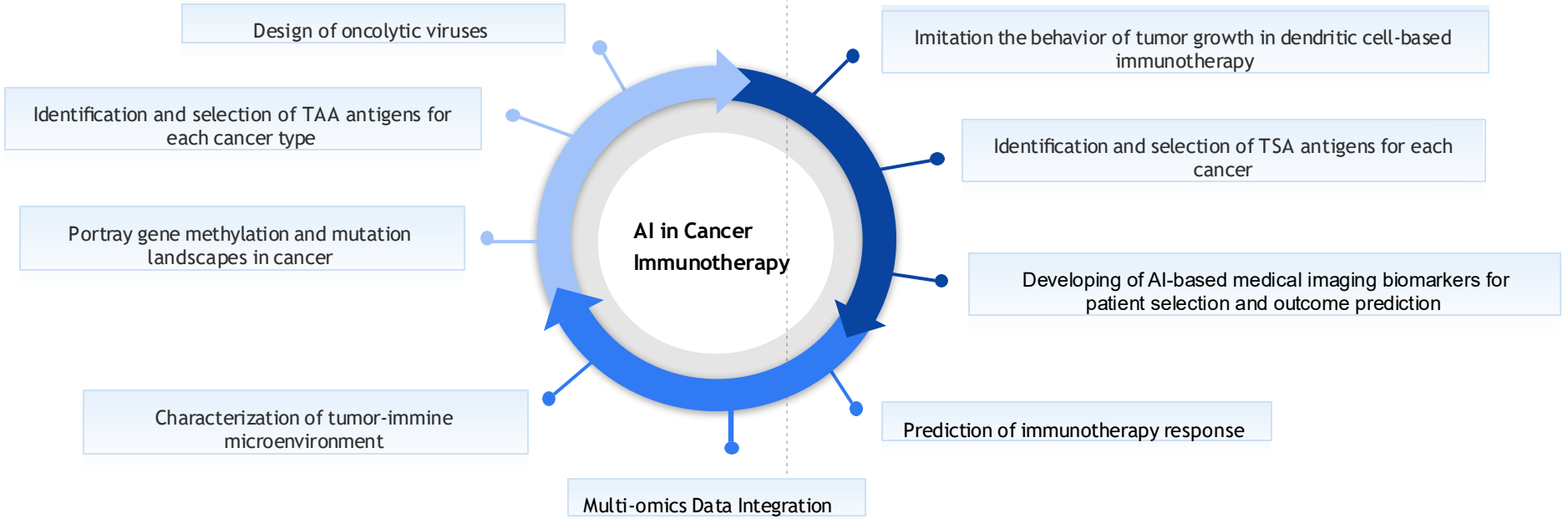
ACCESS | Metrics & More | Article Recommendations | Supporting Information



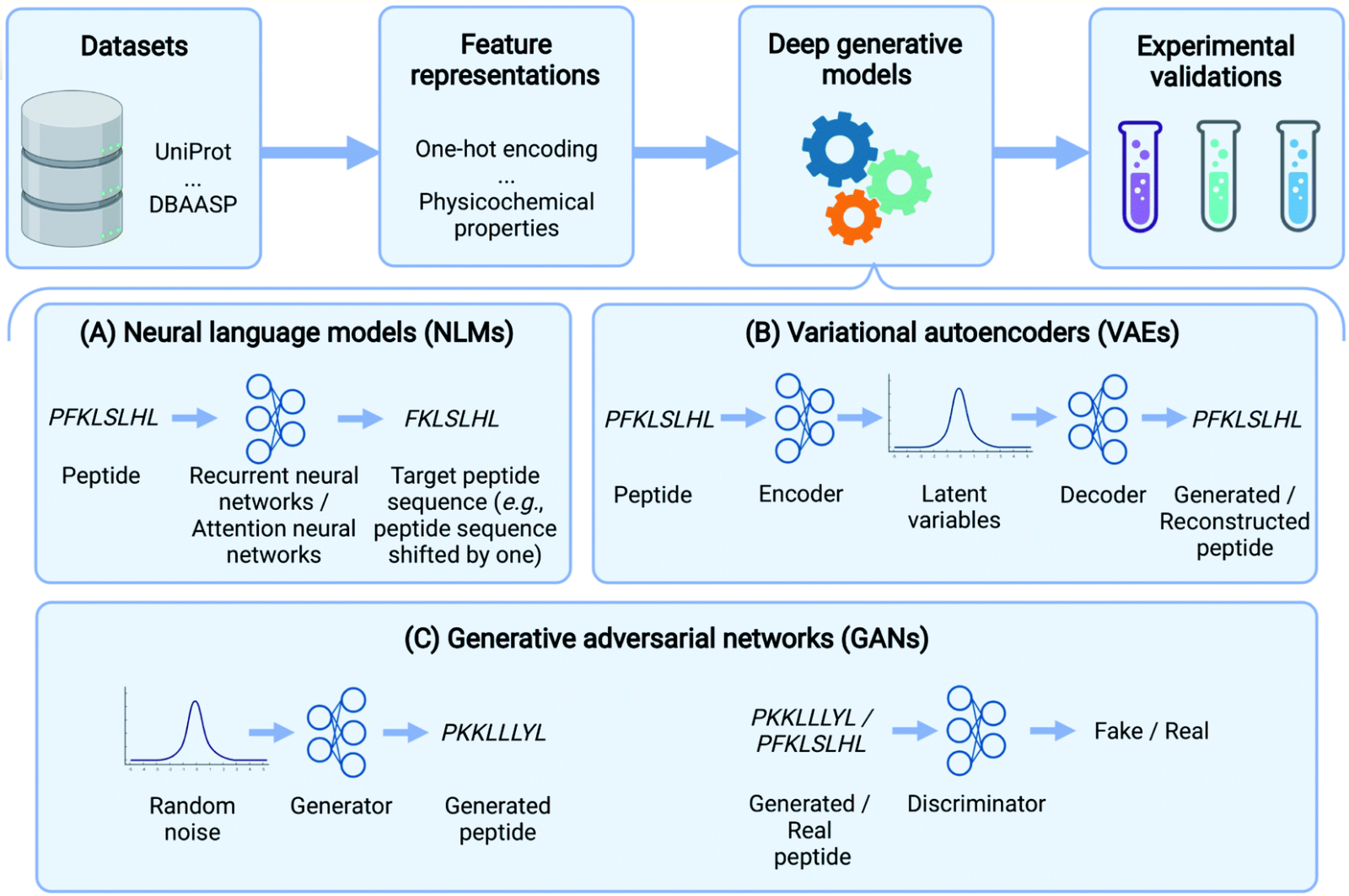

A)
PMO-CPP Library & Activity (600 sequences)
Predictor
Neural Network
Predicted Sequences
Previous Work
Mach miniprotein
40-80 residues
This Work
Short CPPs
<20 residues
Activity
Length
Arginine
K W A X ···
Generator
Neural Network
CPP Thesaurus (>1000 sequences)
Optimizer
Genetic Algorithm
B)
GCT ATT ACC TTA ACC CAG
peptide
Phosphorodiamidate morpholino (PMO)
C)
(1) Internalization
(2) Endosomal Escape
(3) Exon-skipping
(4) EGFP fluorescence



* Source: xxxx, 2022.0x.xx



* Source: xxxx, 2022.0x.xx



* Source: xxxx, 2022.0x.xx

실습

- 실습 코드: https://github.com/fourmodern/2025_aidrugdiscovery
- 구글 코랩: https://colab.research.google.com/
- 구글 AI Studio: https://aistudio.google.com/prompts/new_chat