Deep learning of Protein Sequence Design of Protein-protein Interactions

Raulia Syrlybaeva¹, Eva-Maria Strauch^{1,2}

¹ *Department of Pharmaceutical and Biomedical Sciences, University of Georgia, Athens, GA 30602*

² *Institute of Bioinformatics, University of Georgia, Athens, GA 30602.*

Abstract

Motivation

As more data of experimentally determined protein structures is becoming available, data-driven models to describe protein sequence-structure relationship become more feasible. Within this space, the amino acid sequence design of protein-protein interactions has still been a rather challenging sub-problem with very low success rates - yet it is central for the most biological processes.

Results

We developed an attention-based deep learning model inspired by algorithms used for image-caption assignments for sequence design of peptides or protein fragments. These interaction fragments are derived from and represent core parts of protein-protein interfaces. Our trained model allows the one-sided design of a given protein fragment which can be applicable for the redesign of protein-interfaces or the *de novo* design of new interactions fragments. Here we demonstrate its potential by recapitulating naturally occurring protein-protein interactions including antibody-antigen complexes. The designed interfaces capture essential native interactions with high prediction accuracy and have native-like binding affinities. It further does not need precise backbone location, making it an attractive tool for working with *de novo* design of protein-protein interactions.

Availability

The source code of the method is available at <https://github.com/strauchlab/iNNterfaceDesign>

Supplementary information

Supplementary data are available at Bioinformatics online.

1 Introduction

The ability to computationally engineer protein sequences has a wide range of applications from therapeutics (Fosgerau & Hoffmann, 2015; Khera & Maity, 2019), to vaccines (Li & Li, 2020; Liu et al., 2020; Malonis et al., 2020; X. Zhou et al., 2020), sensors (Karimzadeh et al., 2018; Merkx et al., 2019) or protein-based materials (Capezza et al., 2019; de la Rica & Matsui, 2010) and beyond that. While there has been progress towards designing protein folds, much improvement is needed for the redesign or *de novo* design of protein-protein interfaces (PPIs). The success rates for the *de novo* generation of protein-protein interactions archived by existing methods is very low, with only a few examples demonstrating it is possible (Cao et al., 2020; Fleishman et al., 2011; Strauch et al., 2014). Even most recent work illustrates an average success rate of 0.25% while requiring substantial computational and laboratory resources (Cao et al., 2021) underlining that it is still highly challenging. With DeepMind using deep learning algorithms and entering the field of protein structure prediction, much has changed about how machine learning approaches can be used for protein folding and also design. Recent work using neural networks improved substantially accuracy in structure prediction (Baek et al., 2021; Jumper et al., 2021; Senior et al., 2020) and protein sequence design (Anand et al., 2020; Chen et al., 2020; Gao et al., 2020; O'Connell et al., 2018). The latter methods outperform traditional methods for sequence design based on energy function integrated into sampling, filtering, and optimization procedures (Adolf-Bryfogle et al., 2018; Desjarlais & Handel, 1995; Raha et al., 2000). The

average sequence recovery achieved by the current top-performing protein-design programs is around 30% (J. Zhou et al., 2020), while SPROF model (Chen et al., 2020) achieved 39.8% on independent test sets. Based on these inspiring results, we developed a deep learning-based approach for the sequence design of PPIs which is a sub-problem within the sequence design space.

A main challenge for design of new protein interfaces is the identification and placement of energetic “hot-spots” (Cukuroglu et al., 2014; Wells & Clackson, 1995). A characteristic feature of protein interactions is that interface residues do not contribute equally to the binding energy and only a few residues have energetically highly favorable interactions. Their placement is complicated by the diversity a given rotamer can have or the conformational states a protein can exhibit. Further, models should consider both surface contacts with the target protein, but also protein stability of the unbound state. Here, we report on the recovery of small patches of PPIs containing hot-spot residues. We focus on design of amino acid sequences (AAS) of “sixmer” (six residue long) fragments of protein ligands exhibiting high affinity PPIs. Our method is intended to be used for the recovery of native AAS for a fixed backbone of protein ligand or the design of *de novo* interactions. For the latter case, often, the precise location of the backbone of the ligand is not exactly known which may also be applicable for the re-design of complexes that have been generated through homology modeling. In these cases, deviations in bond lengths, valent or dihedral angles may occur and need to be addressed. To mimic this challenge, all ligands were perturbed before the evaluation of our sequence design model.

Our method differs from existing deep learning models for design of PPIs (Wu et al., 2021; Zhang et al., 2021), which are intended to help docking by defining interaction pairs of residues of two counterparts rather than creating the counterparts itself. The architecture of neural network is inspired and based on a model generating captions of an image with visual attention (Xu et al., 2015) turning pictures into captions. Hence, we treat the structures of protein as the 3D object to be captured. Features are then extracted using machine learning vision techniques and then translate these into sequences specifically focusing on the properties of how its counterpart (protein ligand) should look like. We developed two versions of the deep learning models, named PepSeP1 and PepSeP6, producing single and multiple outputs correspondingly. The methods significantly outperform Rosetta’s FastDesign mover (Khatib et al., 2011; Tyka et al., 2011) on an independent test set. Comparison of binding affinities of native and designed peptides has shown small difference of 0.7 Rosetta energy unit (REU) in favor of native structures according to estimation by means of the current Rosetta scoring function (ref2015,(Alford et al., 2017). As mentioned, the developed model is intended to be applied on difficult cases involving backbones even far from native geometries for which most current methods are not very adapted for. An extension of the method allows the design of larger fragments by making subsequent designs of connected backbone fragments. Considering these results, we provide a highly promising neural network (NN)-based approach for the design of PPIs.

2 Materials and methods

2.1 Test Sets

The method is trained and tested on peptide-binding site complexes, extracted from native PPIs (Fig. 1). The peptide of the complex is 6-residue fragment of a protein ligand and the binding site is a patch of a receptor consisting of 24-48 residues which are closest to backbone atoms of the 6-residue fragment. All complexes for this study originate from multichain structures obtained from Protein Data Bank (Berman, 2000).

The complexes were extracted from 9,101 PDB-files. 70 Files containing complexes of influenza’s hemagglutinin (HA) and coronaviruses MERS-CoV, SARS-CoV and SARS-CoV-2 proteins with antibodies (Table S2) were separated from the main set to form a benchmark set, all other complexes with the listed antigens were deleted from the main set. The files of the main set were randomly split to construct training, validation and test sets in quantities of 8 581, 271, 179 respectively; total number of complexes obtained from these crystal structures are 95 087, 2 933 and 1 261; part of the curation was to eliminate any duplicates of the complexes to enable independent training and test sets. The benchmark set, omitting duplicates, consists of 915 complexes which are independent of the main test set regarding antigens; it contains antibody or antibody fragments which can be present in other datasets.

Our method was evaluated using both test and benchmark sets, referred as “set T” and “set B” respectively. Set T is divided into two subsets T-ho and T-he, based on whether they were derived from homo- and hetero-oligomeric PPIs correspondingly. Set B was used for evaluation of sequence recovery of interfaces focusing on antibody-

antigen protein complexes. This set is divided into subsets as well (Table 1). Quantity of complexes in the benchmark subsets depending on types of antigens is summarized in Table S3. Other details related to construction of the datasets can be found in methods section of SI (SI 1.1 – 1.3).

Figure 1. Generation of datasets of peptide-binding site complexes and overview of structures utilized in the study. **(a)** Crystal structure of protein-protein complex with selected interacting 6-residue fragment of protein ligand. **(b)** Peptide-binding site complex of datasets. **(c)** Types of peptide ligands and their generation: native peptide ligand which is the 6-residue fragment depicted in the subfigure **a** (c1), was mutated into an all-glycine peptide ligand and perturbed (c2, SI 1.2). Perturbed backbones were then either reverted to their native amino acid sequence (c3, and annotated as native* within main text to highlight the backbone perturbation) or designed with PepSePl (c4). HP (highly perturbed) backbones (c5) have non-native backbone geometry after extensive perturbation. s

Table 1. Subsets of the benchmark set

Subset	Source of a backbone fragment (peptide) to be designed	Source of a binding site to which the peptide is attached	Number of complexes
B-ab/ag	Antibody	Antigen	150
B-ag/ab	Antigen	Antibody	144
B-ab/ab	Antibody	Antibody	485
B-ag/ag	Antigen	Antigen	136

Generation of highly perturbed (HP) peptide fragments

Beside testing the methods on all-glycine peptide ligands extracted from native interactions, we tested our method on artificially highly perturbed (HP) peptide variations of a 6-residue all-glycine peptide ligands fragment (Fig. S3). These backbone variations were generated by PepBB deep learning model, developed by our group and soon to be published (available at (*INNterfaceDesign*, 2021)), as a starting point of the complexes of the test set T. Twelve peptide ligands were generated for each of the binding sites, the ones with the lowest RMSD to native peptide ligands were selected as HP peptides. Average RMSD of them is about 5.0 Å (3.8 Å for non-terminal residues). High RMS difference are due to 20% of HP peptides with a reverse orientation to the native conformation: they are on same location but N- and C- termini are other way around (Figure S3a). Terminal residues of the native peptide ligands are typically non-interacting (SI 1.2) which also contributed to RMSD. Due to some deviations from native

backbones in terms of subtle geometrical features, they provide very challenging targets for AAS design and recovery of hot-spot interactions.

2. 2 Input data for deep learning models

Input for the neural networks are geometrical and sequence features of the binding sites and XYZ coordinates of backbone atoms of the peptide ligands. We utilized two versions of the binding sites: first is based on all initially selected residues (24-48), and a second reduced representation, consisting of 20 residues of the binding site closest to the peptide ligands for improved accuracy. As a geometrical feature description of the structures, we used distance maps with either N or O backbone atoms. For each of these two atom types, we generated an intra- (map 1) and intermolecular distance map (map 2), describing either distances for the interface atoms or across the interface respectively. Input arrays and the corresponding descriptors are summarized in Table 2 (details under SI 1.4). Input 5, not mentioned in Table 2, contains binary values (0 or 1) denoting the type of PPI (homo- or hetero-oligomeric).

Table 2. Input from structures used for feature generation for deep learning models during training and testing.

Descriptor		Binding site	Peptide	Reduced binding site
Topological	Distance maps 1 (intramolecular)		Input 4	Input 9
	Distance maps 2 (intermolecular)	Input 1		Input 7
	Secondary Structure	Input 3		Input 8
AAS		Input 2	Input 6	

2.3 Architecture of Developed Neural Networks

For our sequence design algorithm, we developed a two-directional architecture involving two neural networks. The main block produces AAS for 6-residue peptide backbones (output 1) based on input 1-5 (containing AAS of the binding site and topological features of the interfaces, Table 2 and SI 1.4). To improve its accuracy, we developed the second module (reverse prediction block) trained on input 5-9, which will produce output 2, a reduced representation of the target binding site. Hence, output 2 should mirror most of input 2. Training conditions required not only high rates of recovery of native AAS of the peptide, but it also required high accuracy for output 2 when using output 1 in the reverse prediction block as input, instead of input 6 (Fig. 2). The combination of these two neural networks improved sequence recovery and also resulted in more favorable binding energies (> 4 REU improvement) making our model comparable to native interactions. After training, the reverse prediction block becomes redundant, whereas the trained main block will be used for predictions of PepSeP1.

Figure 2 Peptide ligand sequence recovery method by PepSeP1 attention neural network. There are 9 types of input arrays provided to the neural networks (see details under 2.2 and SI 1.4).

Both blocks have an encoder-decoder architecture which has been successfully utilized for sequence prediction models (Chen et al., 2020). The architecture of the blocks is very similar and loosely based on an TensorFlow tutorial (Image Captioning with Visual Attention, n.d.) about how to implement models generating image captions with visual attention (Xu et al., 2015), specifically the decoder part. The encoders are based on convolutional layers.

The PepSeP6 variation of our algorithm, producing six predicted AAS, uses the same infrastructure as PepSeP1. The scheme of PepSeP6 is presented in Fig. S4. Five outputs are generated by passing feature vectors produced by the encoder of the main block of PepSeP1 model to the decoder of PepSeP6 five times; each of five iterations is accompanied by a final hidden state from the previous prediction of the sequence. The sixth sequence is the output of the PepSeP1 model, which is provided to PepSeP6. A single peptide ligand sequence predicted by PepSeP6 is randomly selected from the resulting sequences after and forwarded to the reverse prediction block of PepSeP1 providing the bidirectional approach. The neural network was built in TensorFlow using Keras API (more details under SI 1.4 -1.6).

2.4 Assessment of experimental and predicted peptide ligands

Refinement and calculation of binding energies of complexes of test sets

The target binding sites were relaxed without peptide ligands. Perturbed peptides were then added back either with their native or redesigned AAS (Fig.1). Optimization of side chain conformations of all residues of the complex and subsequent refinement of poses of the peptide ligands was done applying the FastRelax mover three times over 300 step while applying harmonic constraints (standard deviation SD = 1.0) for peptide ligands based on perturbed native-like backbones (c2). Less restrictive constraints (SD=3.0 with width parameter of 2.0 Å) were used for HP peptides (c5) predictions. Binding free energies of the complexes were estimated using InterfaceAnalyzerMover (Stranges & Kuhlman, 2013) with repacking chains after separation. All stages of the refinement were carried out three times using FastRelax mover over 300 steps. The structure with the lowest score out of three results was selected for the next operation. The Rosetta scoring function ref15 was used for all calculations.

Redesign of complexes

Peptide ligands designed by PepSeP1 method underwent an additional design step after refinement using the FastDesign protocol to compare results with original performance of PepSeP1. We set constraints on residue types according to position-specific scoring matrices (PSSM) based on outputs of PepSeP1 (SI 1.7). We performed three different protocols of redesign. We controlled how many possible amino acid types are allowed at each position. Variations are denoted as RD3, RD5 and RD20, reflecting whether either the most probably three or five amino acid types should be used or whether all amino acid were allowed. Probabilities were applied in form of described PSSMs extracted from the PepSeP1 model. In comparison, we also performed redesign using the FastDesign protocol on all-glycine peptide backbones. Two types of relax scripts were utilized during the redesigns: default (MonomerRelax2019) and InterfaceDesign2019.

3 Results

Table 3 summarizes performances of the developed deep learning models alongside with results of different approaches based on FastDesign protocol (Maguire et al., 2021; Tyka et al., 2011) on recovery of native sequences of the peptide ligands of the test and benchmark sets.

Table 3. Rates of recovery of native residues of peptide ligands of sets T and B achieved by PepSeP1 and PepSeP6 methods based on all-glycine peptides, obtained after redesigning of PepSeP1 sequences by means of Rosetta's

FastDesign protocol according to RD3, RD5, RD20 schemes and obtained by FastDesign protocol (MonomerRelax2019)) applied to all-glycine peptide fragments.

Subset	R_{all} , %						$R_{hot-spot}$, %					
	PepSeP1	RD3	RD5	RD20	PepSeP6	FastDesign	PepSeP1	RD3	RD5	RD20	PepSeP6	FastDesign
T-ho	51.55	41.87	36.53	29.73	45.61	23.35	57.75	56.23	55.18	50.41	51.09	44.18
T-he	28.56	27.22	26.1	22.64	26.02	24.93	37.19	41.81	44.01	40.34	31.7	46.67
T	44.42	37.34	33.31	27.54	39.54	23.84	51.07	51.58	51.58	47.16	44.79	44.98
B-ab/ag	28.11	22.98	23.99	23.11	24.7	22.7	33.52	31.41	32.05	33.33	30.72	31.74
B-ag/ab	18.06	20.95	21.43	21.43	16.22	23.24	24.14	33.63	30.97	45.13	19.37	48.67
B-ag/ag	56.25	42.03	37.87	28.43	47.75	22.43	75.52	69.27	64.58	55.73	62.44	40.62
B-ab/ab	65.53	54.7	47.65	37.61	61.4	25.14	77.16	73.89	70.87	65.01	72.69	50.43

Sequence recovery accuracy (R_{all}) on set T is 44.42%. The results depend greatly on types of PPIs: native sequence recovery rates for homo- and hetero-oligomeric PPIs are 51.55% and 28.56%, respectively. Better performance of the method on obligate PPIs is affirmed by accuracy exceeding 56% archived on subsets B-ab/ab and B-ag/ag. Lower rates archived on hetero-oligomeric PPIs are likely due to fact that substantial part of heteromeric protein interfaces are transient interactions and thereby weaker in nature than assembled complexes. Further, residue composition of transient interactions is more diverse and include higher rates of polar and charged groups alongside hydrophobic amino acids (Acuner Ozbabacan et al., 2011). Ligand residues of antigen-antibody interfaces of subsets B-ab/ag and B-ag/ab, which are transient PPIs, are predicted with 28.11% and 18.08% success rate. Antigen structures are represented as a single chain or as a fragment of the chain in the original crystal structures often. Usage of the method on the benchmark set where all hemagglutinin complexes are extracted from trimer structures (the coronavirus complexes are same as in the main benchmark set) provide overall accuracy for subset B-ab/ag of 31.85 %. Detailed data regarding the accuracy of the method depending on types of antigens can be found in Table S4, best results for PepSeP1 are observed in case of SARS-CoV-2: 41.67% and 50% in subsets B-ab/ag and B-ag/ab, correspondingly. Secondary structure of the peptide fragment impacts the accuracy as well (Table S5): the highest recovery rate of 47.11% was observed for β -sheet structures, accuracies of predictions for α -helices and especially for loops were lower (46.17 and 39.35%, respectively).

The performance of the model on recovery of hot-spot residues of the peptides was considered as well (Table 3). Energy contributions of individual residues to the binding ΔG_i were obtained by alanine scanning (Kortemme et al., 2004). Contacts were treated as hot-spots if they had an binding energy contribution of at least 3 REU. Success rate $R_{hot-spot}$ was calculated with respect to hot-spot residues of native peptide ligands. $R_{hot-spot}$ exceeds R_{all} by 6% approximately and equals to 51.07% on set T. The performance of the method is 1.5 times better on hot-spot residues of homo-oligomeric PPIs than of hetero-oligomeric PPIs: corresponding values of $R_{hot-spot}$ are 57.75% and 37.19%. Recovery of hot-spot residues of subsets B-ab/ag and B-ag/ab is worse than on subset T-he. We got the lowest results on subset B-ag/ab, in large part due to poor recovery rates obtained on influenza's hemagglutinin (16.95%, Table S5) which constitute 82% of complexes of this subset. At the same time, $R_{hot-spot}$ observed on SARS-CoV-2 structures of subset B-ag/ab is equal to 100% (Table S5).

Assessing binding free energies $\Delta G_B^{native*}$ and $\Delta G_B^{designed}$ across test subsets, the binding sites have higher affinity for peptide ligands with native sequences binding energies on average, but the difference is small or equal to 0.7 REU on set T only (Table S6). We estimated complexes with native* peptides (see SI 1.2) instead of native ones during calculations of energetic metrics, for comparison with the designed complexes, in order to get rid of systematic superiority of native peptides due to more favorable backbone conformations which did not undergo perturbation. Comparison of native and designed by PepSeP1 method complexes in more details are discussed in sections SI 2.1-2.2. The accuracy of the predictions depends on relative solvent accessible surface area (SASA) of a given residue and $\Delta^i G$ P-value of the native PPIs (Fig. S9). We observed some reduction of R_{all} with increasing of both SASA, as seen before (Chen et al., 2020), and $\Delta^i G$ P-value. As expected, the decrease is more prominent for hetero-oligomeric PPIs. Results of application of PepSeP1 on concrete examples, including design of AAS for antibody fragments bound to SARS-CoV-2 and influenza's HA are presented in SI 2.3.

As there can be different solutions to binding to the same interface (DeLano et al., 2000) we also integrated a variation of our software that produces six sequences, called PepSeP6. The output sequences differ on average

within three positions usually. However, we also saw complete identical sequences or only 1 variation (Fig. S13). Such convergence is observed when all six outputs are generated with high rates of sequence recovery (80-100%).

R_{all} of the model across all six outputs is 39.54% (Table 3), which is lower than R_{all} of PepSeP1. However, R_{all} calculated only across the outputs most matching the native sequences out of six is equal to 52.8% (Table S9). The difference in binding energies of complexes with all six sequences is only within 1 REU in average (Fig. S14), the values are very close to native ones. Thus, PepSeP6 sequences provide comparable binding energies and could present alternatively binding solutions as seen in nature. More discussion of performance of PepSeP6 model is presented in SI 2.4.

The Rosetta FastDesign protocol is currently a commonly used protocols for the redesign of AAS demonstrating many experimentally validated designed protein interfaces (Cao et al., 2020; Huang et al., 2016; Jacobs et al., 2016; Linsky et al., 2020; Silva et al., 2019). Here, we compared the performance PepSeP1 model with FastDesign itself and with results of a combination these two methods in three versions, called RD3, RD5 and RD20, described in methods section. FastDesign results obtained using the default relax script are shown here (Table 3). Overall, redesigns of PepSeP1 sequences did not result in higher rates of recovery of native residues according to R_{all} values of RD3, RD5 and RD20 approaches. However, redesigns provide slightly higher rates $R_{hot-spot}$ in comparison with PepSeP1 in case of hetero-oligomeric PPIs. PepSeP1 substantially outperformed FastDesign in sequence recovery when both are applied to the same all-glycine peptides. Rates of recovery obtained using InterfaceDesign2019 relax script are reported in Table S10.

HP peptide ligands having large RMSD values and indistinct subtle geometrical features which might have been used to identify residues (like dihedral angles characteristic for proline), but overall relevant for target binding sites nevertheless (Fig. S3), are very challenging targets for AAS design and recovery of hot-spot interactions comparatively to all-glycine peptides based on the native backbones. Thus, FastDesign incorporates an excessive number of glycine and proline residues to sequences for HP peptides (Fig. S16), the average binding energy is about -7 REU (Fig. S15) only, while ΔG_B^{native} is -17.1 REU. Therefore, we developed PepSeP1 originally to work in conjunction with PepBB or on other peptide backbones engendered artificially for the *de novo* design of PPIs and having deviations from native backbones of different extents. The distribution of amino acid types in sequences predicted by PepSeP1 resembles the native distribution far better than FastDesign (Fig. S16), the average binding energy is -15.1 REU. Therefore, the performance of PepSeP1 on non-native peptide ligand fragments is far more relevant in comparison with FastDesign.

Figure 3. Design of highly perturbed (HP) peptide fragment generated as a binder for influenza's hemagglutinin and located nearby CDR-loop of FI6V3 antibody structure (PDB code: 3ZTJ) by PepSeP6, FastDesign and RD20 methods. ΔG_B is in REU.

Figure 3 illustrate application of PepSeP6 method for design of AAS for HP peptide ligand generated for a highly conserved stalk region in HA and located with RMSD of 2.9 Å relative the CDR-loop of FI6 antibody; (the first

design corresponds to PepSep1 output). FI6 antibody fragment has three hot-spots at considered region: Arg99 ($\Delta\Delta G_i = 3.2$ REU), Leu100A ($\Delta\Delta G_i = 4.1$ REU), Tyr100C ($\Delta\Delta G_i = 3.9$ REU). Among PepSep6 outputs, four designs contain recovered Leu100A interaction, two designs have Tyr near C-terminus and two designs have Arg near N-terminus; some of them provide two or more native-like recovered amino acids, as in complex selected for the illustration of the results. At the same time, FastDesign performed much poorer. Redesign of the complex obtained using PepSep6 method through RD20 approach allowed to keep native-like interactions involving Arg and Leu and to add recovered Ser residue.

The results show that PepSep1 and PepSep6 can reproduce relevant contacts; the resulted peptide ligands show high binding affinity and often outperform designs of FastDesign in case of redesigning a highly perturbed native backbones and thereby illustrates its usefulness for homology models as well as or potentially *de novo* designed backbones as they likely are not at the exact position they should be – as either methods have a high margin of error. Thus, a more knowledge-based design process can even guide the backbone refinement. The PepSep6 method provides more diverse designs, iteration through which can reveal sequences with higher recovery rates, but it will be also useful for any affinity optimization processes as there are multiple solution to binding at a specific epitope.

Conclusions

The neural network designed for recovery of peptide ligand sequences at a known protein binding site is performed in this study. To our knowledge, this is the first neural network model for prediction of amino acid sequences for peptides involved into interchain interactions. The model was developed in two versions: PepSep1 and PepSep6 with single and multiple (six) outputs, correspondingly. The native sequence recovery rate of PepSep1 is 44.42% on the independent test set; the average accuracy of PepSep6 designs is 39.54% with recovery rates of 52.8% across the output sequences resembling the native structures the most. The models characterized by intentional training on non-perfect backbones of structures which make them more applicable for non-perfect engineered *de novo* interactions, including homology models of PPI if one of the counterparts is intended for a redesign.

References

- Acuner Ozbabacan, S. E., Engin, H. B., Gursoy, A., & Keskin, O. (2011). Transient protein–protein interactions. *Protein Engineering, Design and Selection*, 24(9), 635–648. <https://doi.org/10.1093/protein/gzr025>
- Adolf-Bryfogle, J., Kalyuzhnyi, O., Kubitz, M., Weitzner, B. D., Hu, X., Adachi, Y., Schief, W. R., & Dunbrack, R. L. (2018). *RosettaAntibodyDesign (RAbD) : A general framework for computational antibody design*.
- Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O’Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6), 3031–3048. <https://doi.org/10.1021/acs.jctc.7b00125>
- Anand, N., Eguchi, R. R., Derry, A., Altman, R. B., & Huang, P.-S. (2020). Protein Sequence Design with a Learned Potential. *BioRxiv*. <https://doi.org/10.1101/2020.01.06.895466>
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. <https://doi.org/10.1126/science.abj8754>
- Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242. <https://doi.org/10.1093/nar/28.1.235>
- Cao, L., Coventry, B., Goreshnik, I., Huang, B., Park, J. S., Jude, K. M., Marković, I., Kadam, R. U., Verschueren, K. H. G., Verstraete, K., Walsh, S. T. R., Bennett, N., Phal, A., Yang, A., Kozodoy, L., DeWitt, M., Picton, L., Miller, L., Strauch, E.-M., ... Baker, D. (2021). Robust de novo design of protein binding proteins from target structural information alone. *BioRxiv*. <https://doi.org/10.1101/2021.09.04.459002>
- Cao, L., Goreshnik, I., Coventry, B., Case, J. B., Miller, L., Kozodoy, L., Chen, R. E., Carter, L., Walls, A. C., Park, Y.-J., Strauch, E.-M., Stewart, L., Diamond, M. S., Veessler, D., & Baker, D. (2020). De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. *Science*, 370(6515), 426–431. <https://doi.org/10.1126/science.abd9909>
- Capezza, A. J., Newson, W. R., Olsson, R. T., Hedenqvist, M. S., & Johansson, E. (2019). Advances in the Use of Protein-Based Materials: Toward Sustainable Naturally Sourced Absorbent Materials. *ACS Sustainable Chemistry & Engineering*, 7(5), 4532–4547. <https://doi.org/10.1021/acssuschemeng.8b05400>
- Chen, S., Sun, Z., Liu, Z., Liu, X., Chong, Y., Lu, Y., Zhao, H., & Yang, Y. (2020). To Improve Protein Sequence Profile Prediction through Image Captioning on Pairwise Residue Distance Map. *J. Chem. Inf. Model.*, 60(1), 391–399. <https://doi.org/10.1101/628917>
- Cukuroglu, E., Engin, H. B., Gursoy, A., & Keskin, O. (2014). Hot spots in protein-protein interfaces: Towards drug discovery. *Progress in Biophysics and Molecular Biology*, 116(2–3), 165–173. <https://doi.org/10.1016/j.pbiomolbio.2014.06.003>
- de la Rica, R., & Matsui, H. (2010). Applications of peptide and protein-based materials in bionanotechnology. *Chemical Society Reviews*, 39(9), 3499–3509. <https://doi.org/10.1039/b917574c>
- DeLano, W. L., Ultsch, M. H., De Vos, A. M., & Wells, J. A. (2000). Convergent solutions to binding at a protein-protein interface. *Science*, 287(5456), 1279–1283. <https://doi.org/10.1126/science.287.5456.1279>
- Desjarlais, J. R., & Handel, T. M. (1995). De novo design of the hydrophobic cores of proteins. *Protein Science : A Publication of the Protein Society*, 4(10), 2006–2018. <https://doi.org/10.1002/pro.5560041006>
- Fleishman, S. J., Whitehead, T. A., Ekiert, D. C., Dreyfus, C., Corn, J. E., Strauch, E.-M., Wilson, I. A., & Baker, D. (2011). Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science (New York, N.Y.)*, 332(6031), 816–821. <https://doi.org/10.1126/science.1202617>
- Fosgerau, K., & Hoffmann, T. (2015). Peptide therapeutics: current status and future directions. *Drug Discovery Today*, 20(1), 122–128. <https://doi.org/https://doi.org/10.1016/j.drudis.2014.10.003>
- Gao, W., Mahajan, S. P., Sulam, J., & Gray, J. J. (2020). Deep Learning in Protein Structural Modeling and Design. *Patterns*, 1(9), 100142. <https://doi.org/10.1016/j.patter.2020.100142>
- Huang, P.-S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, 537(7620), 320–327. <https://doi.org/10.1038/nature19946>
- Image captioning with visual attention*. (n.d.). https://www.tensorflow.org/tutorials/text/image_captioning

- iNNterfaceDesign*. (2021). <https://github.com/strauchlab/iNNterfaceDesign>
- Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., & Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. *Science*, 352(6286), 687–690. <https://doi.org/10.1126/science.aad8036>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Karimzadeh, A., Hasanzadeh, M., Shadjou, N., & de la Guardia, M. (2018). Peptide based biosensors. *TrAC Trends in Analytical Chemistry*, 107, 1–20. <https://doi.org/https://doi.org/10.1016/j.trac.2018.07.018>
- Khatib, F., Cooper, S., Tyka, M. D., Xu, K., Makedon, I., Popović, Z., Baker, D., & Players, F. (2011). Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences*, 108(47), 18949–18953. <https://doi.org/10.1073/pnas.1115898108>
- Khera, H. K., & Maity, K. (2019). Rational Design of Next-Generation Therapeutic Antibodies Using Protein Engineering Tools. In *Next Generation Biomanufacturing Technologies* (Vol. 1329, pp. 109-139 SE – 6). American Chemical Society. <https://doi.org/doi:10.1021/bk-2019-1329.ch006>
- Kortemme, T., Kim, D. E., & Baker, D. (2004). Computational alanine scanning of protein-protein interfaces. *Science's STKE: Signal Transduction Knowledge Environment*, 2004(219), pl2. <https://doi.org/10.1126/stke.2192004pl2>
- Li, W.-H., & Li, Y.-M. (2020). Chemical Strategies to Boost Cancer Vaccines. *Chemical Reviews*, 120(20), 11420–11478. <https://doi.org/10.1021/acs.chemrev.9b00833>
- Linsky, T. W., Vergara, R., Codina, N., Nelson, J. W., Walker, M. J., Su, W., Barnes, C. O., Hsiang, T.-Y., Esser-Nobis, K., Yu, K., Reneer, Z. B., Hou, Y. J., Priya, T., Mitsumoto, M., Pong, A., Lau, U. Y., Mason, M. L., Chen, J., Chen, A., ... Silva, D.-A. (2020). De novo design of potent and resilient hACE2 decoys to neutralize SARS-CoV-2. *Science*, 370(6521), 1208–1214. <https://doi.org/10.1126/science.abe0075>
- Liu, Y., Wang, K., Massoud, T. F., & Paulmurugan, R. (2020). SARS-CoV-2 Vaccine Development: An Overview and Perspectives. *ACS Pharmacology & Translational Science*, 3(5), 844–858. <https://doi.org/10.1021/acspstsci.0c00109>
- Maguire, J. B., Haddox, H. K., Strickland, D., Halabiya, S. F., Coventry, B., Griffin, J. R., Pulavarti, S. V. S. R. K., Cummins, M., Thieker, D. F., Klavins, E., Szyperski, T., DiMaio, F., Baker, D., & Kuhlman, B. (2021). Perturbing the energy landscape for improved packing during computational protein design. *Proteins: Structure, Function, and Bioinformatics*, 89(4), 436–449. <https://doi.org/https://doi.org/10.1002/prot.26030>
- Malonis, R. J., Lai, J. R., & Vergnolle, O. (2020). Peptide-Based Vaccines: Current Progress and Future Challenges. *Chemical Reviews*, 120(6), 3210–3229. <https://doi.org/10.1021/acs.chemrev.9b00472>
- Merkx, M., Smith, B., & Jewett, M. (2019). Engineering Sensor Proteins. *ACS Sensors*, 4(12), 3089–3091. <https://doi.org/10.1021/acssensors.9b02459>
- O'Connell, J., Li, Z., Hanson, J., Heffernan, R., Lyons, J., Paliwal, K., Dehzangi, A., Yang, Y., & Zhou, Y. (2018). SPIN2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins*, 86(6), 629–633. <https://doi.org/10.1002/prot.25489>
- Raha, K., Wollacott, A. M., Italia, M. J., & Desjarlais, J. R. (2000). Prediction of amino acid sequence from structure. *Protein Science*, 9(6), 1106–1119. <https://doi.org/10.1110/ps.9.6.1106>
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710. <https://doi.org/10.1038/s41586-019-1923-7>
- Silva, D.-A., Yu, S., Ulge, U. Y., Spangler, J. B., Jude, K. M., Labão-Almeida, C., Ali, L. R., Quijano-Rubio, A., Ruterbusch, M., Leung, I., Biary, T., Crowley, S. J., Marcos, E., Walkey, C. D., Weitzner, B. D., Pardo-Avila, F., Castellanos, J., Carter, L., Stewart, L., ... Baker, D. (2019). De novo design of potent and selective mimics of IL-2 and IL-15. *Nature*, 565(7738), 186–191. <https://doi.org/10.1038/s41586-018-0830-7>
- Stranges, P. B., & Kuhlman, B. (2013). *A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds*. 22, 74–82. <https://doi.org/10.1002/pro.2187>

- Strauch, E.-M., Fleishman, S. J., & Baker, D. (2014). Computational design of a pH-sensitive IgG binding protein. *Proceedings of the National Academy of Sciences*, 111(2), 675–680. <https://doi.org/10.1073/pnas.1313605111>
- Tyka, M. D., Keedy, D. A., André, I., Dimaio, F., Song, Y., Richardson, D. C., Richardson, J. S., & Baker, D. (2011). Alternate states of proteins revealed by detailed energy landscape mapping. *Journal of Molecular Biology*, 405(2), 607–618. <https://doi.org/10.1016/j.jmb.2010.11.008>
- Wells, J. A., & Clackson, T. (1995). A hot spot of binding energy in a hormone-receptor interface. *Science*, 267(July), 383–386.
- Wu, T., Hou, J., Adhikari, B., & Cheng, J. (2021). Analysis of several key factors influencing deep learning-based inter-residue contact prediction. *Bioinformatics*, 36(4), 1091–1098. <https://doi.org/10.1093/bioinformatics/btz679>
- Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *32nd International Conference on Machine Learning, ICML 2015*, 3, 2048–2057.
- Zhang, H., Bei, Z., Xi, W., Hao, M., Ju, Z., Saravanan, K. M., Zhang, H., Guo, N., & Wei, Y. (2021). Evaluation of residue-residue contact prediction methods: From retrospective to prospective. In *PLoS Computational Biology* (Vol. 17, Issue 5). <https://doi.org/10.1371/journal.pcbi.1009027>
- Zhou, J., Panaitiu, A. E., & Grigoryan, G. (2020). A general-purpose protein design framework based on mining sequence-structure relationships in known protein structures. *Proceedings of the National Academy of Sciences*, 117(2), 1059–1068. <https://doi.org/10.1073/pnas.1908723117>
- Zhou, X., Jiang, X., Qu, M., Aninwene, G. E., Jucaud, V., Moon, J. J., Gu, Z., Sun, W., & Khademhosseini, A. (2020). Engineering Antiviral Vaccines. *ACS Nano*, 14(10), 12370–12389. <https://doi.org/10.1021/acsnano.0c06109>