nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00803-z

Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning

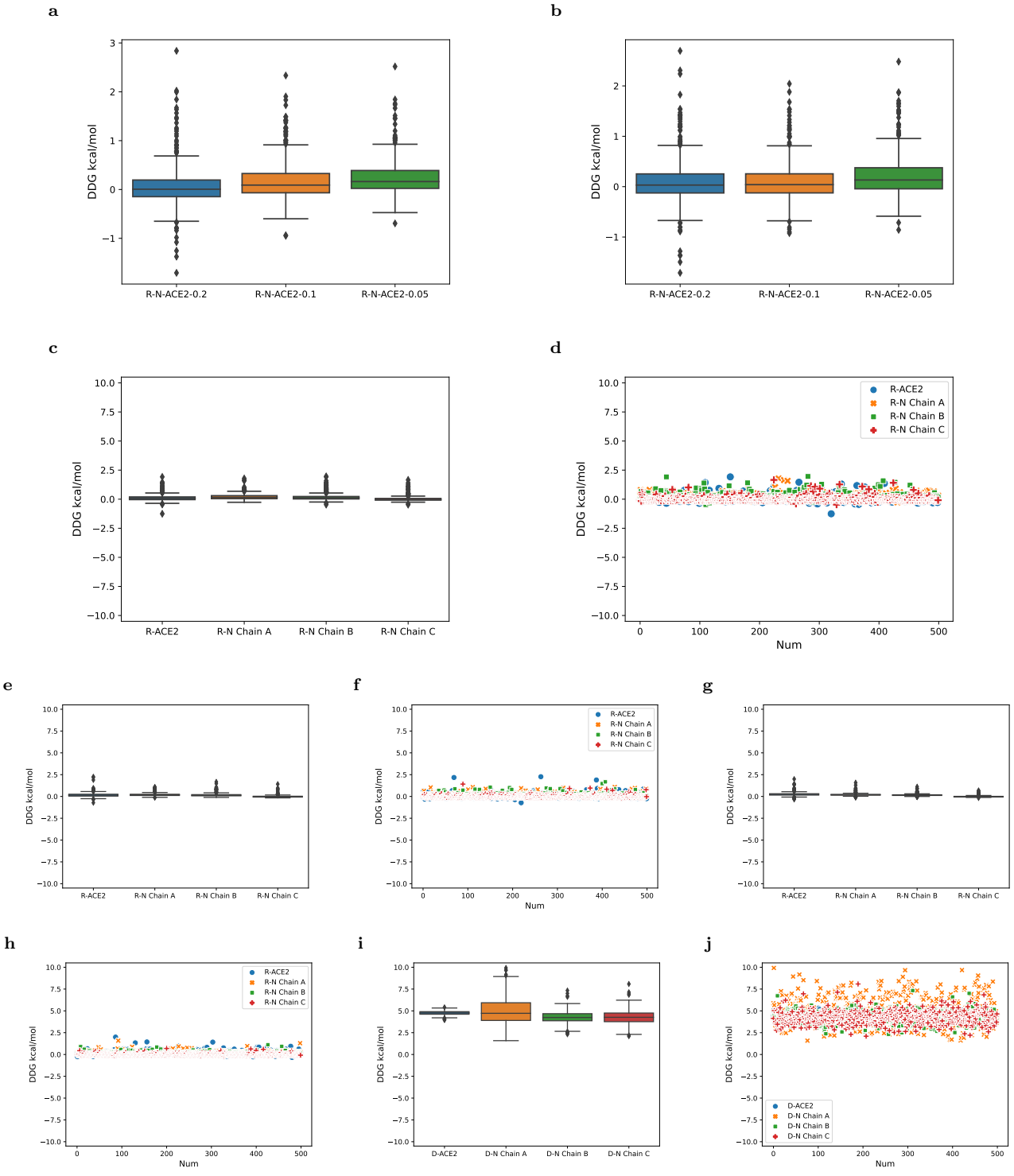
In the format provided by the authors and unedited

The supplementary information includes the following subsections:

1 Supplementary Figures

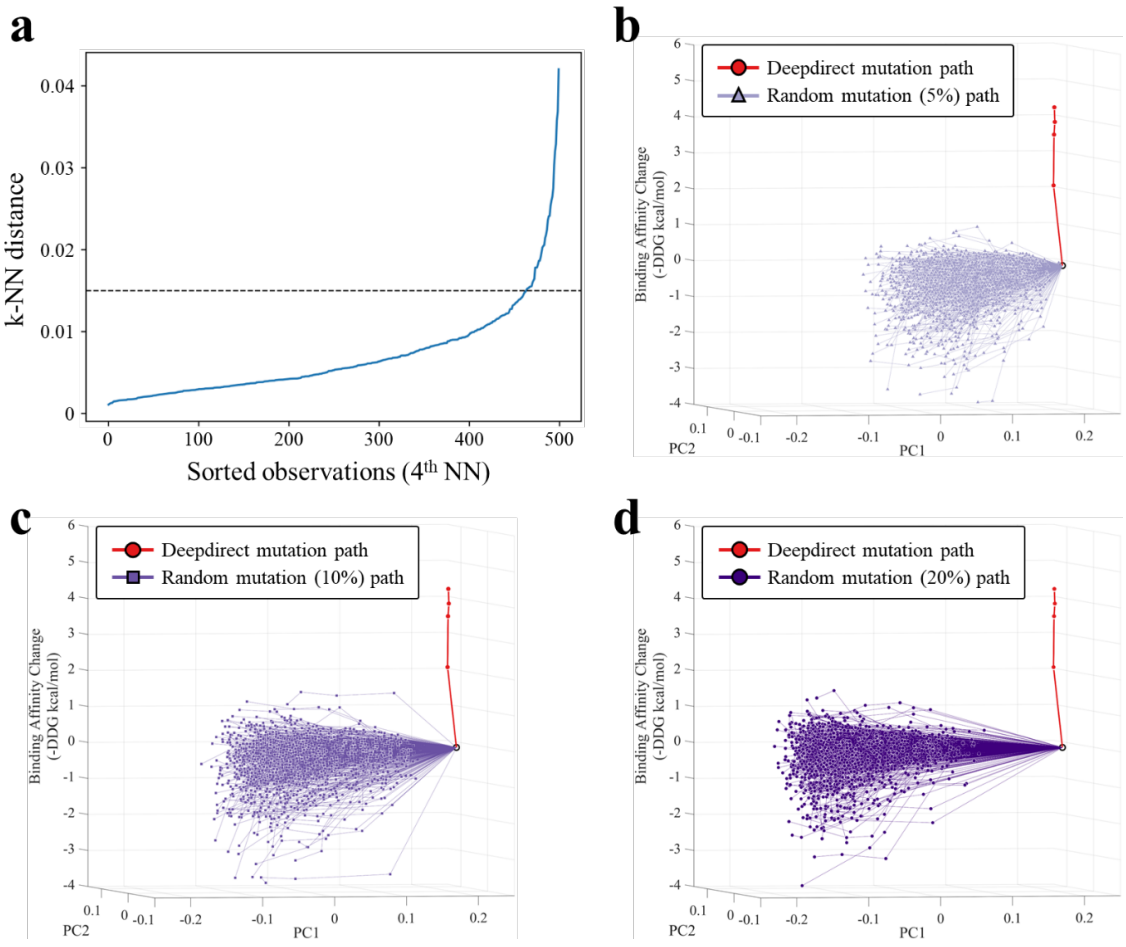
2 Supplementary Notes

3 References



Supplementary Figure 1: **Comparison between Deepdirect- and randomly-generated mutations on Novavax-ACE2 complex.** **(a)**: boxplot plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with different ratio on Novavax-ACE2 complex. n=500 independent samples for each condition. The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **(b)**: boxplot plot of binding free energy change of Novavax-ACE2 complex after randomly-generated mutations with different ratio on Novavax three chains together. n=500 independent samples for each condition. The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **Caption continues on the next page.**

Continued caption of Supplementary Figure 1 from the previous page:
(c): boxplot plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 20% mutation ratio on Novavax-ACE2 complex subset by chain. n=500 independent samples for each condition.The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **(d)**: scatter plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 20% mutation ratio on Novavax-ACE2 complex subset by chain. **(e)**: boxplot plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 10% mutation ratio on Novavax-ACE2 complex subset by chain. n=500 independent samples for each condition. The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **(f)**: scatter plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 10% mutation ratio on Novavax-ACE2 complex subset by chain. **(g)**: boxplot plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 5% mutation ratio on Novavax-ACE2 complex subset by chain. n=500 independent samples for each condition. The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **(h)**: scatter plot of binding free energy change of Novavax-ACE2 after randomly-generated mutations with 5% mutation ratio on Novavax-ACE2 complex subset by chain. **(i)**: boxplot plot of binding free energy change of Novavax-ACE2 after *deepdirect* $model_{dec}$-generated mutations on Novavax-ACE2 complex subset by chain. n=500 independent samples for each condition. The box plots show median, first and third quartiles, minimum and maximum. Outliers are classified as being 1.5 times outside the interquartile range. **(j)**: scatter plot of binding free energy change of Novavax-ACE2 after *deepdirect* $model_{dec}$-generated mutations on Novavax-ACE2 complex subset by chain.



Supplementary Figure 2: **SARS-CoV-2 omicron virus evolution analysis.** **(a)**: KNN distance **(b-d)**: Deepdirect-generated binding affinity-guided SARS-CoV-2 omicron virus spike protein mutation path compare to a batch of 500 randomly generated mutations under mutation ratio of 5%, 10% and 20%.

Supplementary Note 1: Deepdirect potential application scenario

DeepDirect will be useful to optimise various fluorescent antibody structures, which could increase fluorescence detection sensitivity especially in the detection of low-expressed proteins. Similar strategy could be also applied to other conjugated antibodies such as ELISA antibodies, Western-blotting antibodies and so on. Besides, research-used recombinant ligands or signalling molecules could be engineered and optimised by DeepDirect for achieving better research outcomes such as interferons (IFNs), tumor necrosis factors (TNFs), transforming growth factors (TGFs), etc. More importantly, medically used recombinant proteins-based therapies can also be benefited from the optimisation of protein structures by DeepDirect, which is feasible for both public medication and personalised medicine, such as optimising Programmed death-ligand 1 (PDL1) in PDL1 immunotherapy, optimising Chimeric antigen receptor (CAR) in CAR-T cell therapy or optimising IFN in interferon therapy.

Supplementary Note 2: Review of machine learning methods for sequence-to-affinity/sequence-to-structure/function-to-sequence predictions

There are tailored machine learning methods for the prediction of binding affinities between a fixed pair of proteins (i.e., the mutation-to-affinity prediction) for example through regression method PPApred [18] or ISLAND [1]. Other methods DeepDTA [7], Pafnucy [11] and MdeePred [9] have also been developed for the prediction of protein binding affinities but through convolutional neural network architectures together with sequence features and auxiliary information of the chain index. However, these methods are unable to answer our questions: which putative sequence will have a guaranteed binding affinity increase or what mutated sequence will reach a maximum binding affinity with the receptor. Other seminal deep learning algorithms [4, 12, 8, 16, 14, 2, 13] have just aimed to make accurate protein *sequence-to-structure* predictions or protein *function-to-sequence* predictions. In particular, the breakthrough neural network method AlphaFold [10] and its successor Alphafold2 [4] have demonstrated unprecedent accurate prediction capability in the 13th and 14th Critical Assessment of Techniques for Protein Structure Prediction (CASP13 and CASP14). Algorithm ProteinMPNN [3] implements an encoder-decoder architecture to extract protein conformation information and makes the structure-to-sequence prediction. More recently, deep learning frameworks have also been developed for generating non-existing symmetric protein homo-oligomers [15], luciferases [17] and for peptide-binding specificity prediction [6] and protein side-chain packing tasks [5].

Supplementary Note 3: An order of magnitude more DDG change by DeepDirect

We grouped those randomly mutated sequences that have a negative DDG (i.e., binding affinity increased upon the mutation) and those that have a positive DDG (i.e., binding affinity decreased upon mutation). We denote them as D_{RNinc} and D_{RNdec}. Then $D_{RNinc0.05}$ (mutation ratio at 5%), $D_{RNinc0.1}$ (mutation ratio at 10%) and $D_{RNinc0.2}$ (mutation ratio at 20%) have an average DDG -0.12, -0.16 and -0.21 with a median -0.09, -0.12 and -0.16 respectively; while $D_{RNdec0.05}$, $D_{RNdec0.1}$, and $D_{RNdec0.2}$ have an average DDG 0.34, 0.36 and 0.326 with a median 0.24, 0.21 and 0.187. These comparative results demonstrate that DeepDirect's mutation mechanism is very effective to generate a new sequence that has an order of magnitude more binding affinity increase or decrease than the random mutants' affinity change.

Due to the limitation of sample size in the training datasets AB-bind and Skempi2, we were unable to train DeepDirect using mutation data specified by any independent chain. We instead used the data that contains integrated information of the chain mutations to investigate DeepDirect's capability to generate affinity-guided mutations at a specified chain of the protein complex. Therefore, we subset the above generated mutations on each chain of the Novavax-ACE2 complex respectively while keeping the other chains from any mutation to evaluate their mutation effect.

We found that for mutations in ACE2, or in chain A, B or C in Novavax construct, the changes of binding free energy compared to the original complex in all the 500 mutations have a mean -1.996, -3.563, -2.668 or -2.711 with a median -2.038, -3.493, -2.719 or -2.739 respectively (by $model_{inc}$), and a DDG 4.760, 4.996, 4.255 or 4.267 with a median 4.761, 4.721, 4.242 or 4.263 (by $model_{dec}$) respectively (see Fig 4(b) and Supplementary Fig 1(i,j)). In comparison, the changes of binding free energy by randomly generated mutations in ACE2, or in chain A, B or C in the Novavax construct have a mean 0.25, 0.21, 0.17 or 0.00 with a median 0.25, 0.20, 0.15 or -0.02 (under the 5% mutation rate); 0.17, 0.20, 0.17 or 0.00 with a median 0.15, 0.18, 0.13 or -0.03 (under the 10% mutation rate), and 0.102, 0.200, 0.187 or 0.030 with a median 0.070, 0.160, 0.111 or -0.034 (under the 20% mutation rate). These

results suggest that DeepDirect's mutations at each independent chain also follows the specified DDG-changing direction, and the DDG change is again an order of magnitude more significant than the random mutations.

Supplementary Note 4: Further discussion on Deepdirect

DeepDirect predicts the binding affinity change DDG values instead of actual binding affinity of the complex as DeepDirect's main goal is to find a path where the protein mutants undergo incremental changes of binding affinity at a specified direction. In fact, binding affinity changes represent the alteration in the protein's functionality after mutation which is of biological significance.

As we noted in Fig 6(a,b) in the manuscript, DeepDirect converged after four iterations of mutations, reaching a maximum binding affinity for the complex. That is, the binding affinity change will be little or become an opposite direction in the further rounds of mutations. However, the iteration rounds for DeepDirect to monotonically reach the point of maximum binding affinity may differ from one protein complex to another, heavily depending on the sequence properties of the initial proteins as input. It is theoretically difficult to prove exactly at which iteration DeepDirect converges for a protein complex. In algorithm, DeepDirect will always converge within a limited number of iterations for any protein sequences as input. The reason is that we have set a threshold of mutation ratio which prevents an excessive number of base substitutions for the protein.

References

[1] Wajid Arshad Abbasi, Adiba Yaseen, Fahad Ul Hassan, Saiqa Andleeb, and Fayyaz Ul Amir Afsar Minhas. Island: in-silico proteins binding affinity prediction using sequence information. *BioData Mining*, 13(1):1–13, 2020.

[2] Minkyung Baek and David Baker. Deep learning and protein structure modeling. *Nature methods*, 19(1):13–14, 2022.

[3] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022.

[4] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[5] Matthew McPartlon and Jinbo Xu. An end-to-end deep learning method for protein side-chain packing and inverse folding. *Proceedings of the National Academy of Sciences*, 120(23):e2216438120, 2023.

[6] Amir Motmaen, Justas Dauparas, Minkyung Baek, Mohamad H Abedi, David Baker, and Philip Bradley. Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proceedings of the National Academy of Sciences*, 120(9):e2216697120, 2023.

[7] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[8] Nicolas Renaud, Cunliang Geng, Sonja Georgievska, Francesco Ambrosetti, Lars Ridder, Dario F Marzella, Manon F Réau, Alexandre MJJ Bonvin, and Li C Xue. Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nature communications*, 12(1):1–8, 2021.

[9] Ahmet Süreyya Rifaioglu, Rengül Cetin Atalay, D Cansen Kahraman, Tunca Doğan, Maria Martin, and Volkan Atalay. Mdeepred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery. *Bioinformatics*, 37(5):693–704, 2021.

[10] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[11] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

[12] Jérôme Tubiana, Dina Schneidman-Duhovny, and Haim J Wolfson. Scannet: An interpretable geometric deep learning model for structure-based protein binding site prediction. *Nature Methods*, pages 1–10, 2022.

[13] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pages 1–4, 2023.

[14] Michael D Ward, Maxwell I Zimmerman, Artur Meller, Moses Chung, SJ Swamidass, and Gregory R Bowman. Deep learning the structural determinants of protein biochemical properties by comparing structural ensembles with diffnets. *Nature communications*, 12(1):1–12, 2021.

[15] BIM Wicky, LF Milles, A Courbet, RJ Ragotte, J Dauparas, E Kinfu, S Tipps, RD Kibler, M Baek, F DiMaio, et al. Hallucinating symmetric protein assemblies. *Science*, 378(6615):56–61, 2022.

[16] Jinbo Xu, Matthew Mcpartlon, and Jin Li. Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, 3(7):601–609, 2021.

[17] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z Zhang, Ivan Anishchenko, et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023.

[18] K Yugandhar and M Michael Gromiha. Protein–protein binding affinity prediction from amino acid sequence. *Bioinformatics*, 30(24):3583–3589, 2014.