nature portfolio

Corresponding author(s): Habib Bashour and Victor Greiff

Last updated by author(s): Jun 11, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data of the native antibody repertoire was collected from the Observed Antibody Space (OAS: https://opig.stats.ox.ac.uk/webapps/oas/) as well as from primary BCR sequencing data generated using the materials and workflow detailed in the the Methods of this manuscript. The therapeutic monoclonal antibody dataset was downloaded from Thera-SAbDab (https://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/therasabdab/search/). The Humanized mouse (Kymouse) dataset was collected from OAS. The patented antibody database (PAD) was obtained from NaturalAntibody under non-commercial non-disclosure agreement. The validation antibody datasets were downloaded from the AbDb database (http://www.abybank.org/abdb/).
Data analysis	The following programs were used in this study: MiXCR (version 3.0.1), Vagrant VirtualBox (version 2.2.16), SoluProt (version 1.0), netMHCIIpan version 4.0, Reduce (version 3.24.130724), FreeSASA (version 2.1.0), PROPKA (version 3.4.0), ProDy (version 2.0), Arpeggio (version 1.4.1), PyMOL (version 2.5.5). R packages: Peptides (version 2.4.4), ComplexHeatmap (version 2.9.4), factoextra (version 1.0.7), stringdist (version 0.9.8), missRanger (version 2.2.1). Python packages: Levenshtein (version 0.20.8), scikit-learn (version 1.1), BioPython (version 1.79). ABC-EDA algorithm was coded starting from the pseudocode described in the publication cited in this work, and published in this public GitHub repository: https://github.com/csi-greifflab/mwds_calculator The ESM-1v protein language model codes were downloaded from https://huggingface.co/docs/transformers/model_doc/esm

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Our code and low-size data (scripts and figures) are available on GitHub: https://github.com/csi-greifflab/developability_profiling.
Additional structural data, such as predicted models and MD trajectories, are stored on Zenodo: 10.5281/zenodo.10013525.
Raw sequencing data for the experimental BCR sequences were deposited in Sequence Read Archive (BioProject number PRJNA1043047).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex: Male
Reporting on race, ethnicity, or other socially relevant groupings	Race: Caucasian
Population characteristics	One (only) healthy human donor
Recruitment	Voluntary recruitment
Ethics oversight	Sample acquisition was approved by the Regional Ethics Committee of South-Eastern Norway (project 6544)

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For non-native (human-engineered) antibody datasets, the sample size is equal to all antibodies that we could derive from the corresponding public or commercial databases that satisfied the inclusion criteria detailed in the methods section for each of these datasets. For the native antibody dataset, we ensured (1) the augmentation of large-scale (~2.2 Million in total) antibodies in one datasets while keeping the overall computation of developability parameters and antibody high-throughput modeling feasible, and (2) the harmonized inclusion of antibody count per isotype/species combination. For molecular dynamic simulations and their corresponding developability studies, we included five antibodies from the validation dataset (as explained in the methods) to ensure computational feasibility.
Data exclusions	No data exclusion, except for: (1) excluding 0.8% of the total data points (12,500,000) including in the analysis mentioned in Supp Figure 16C. (2) The human-light and all mouse antibodies were excluded from predictability tasks, as the size of the human-heavy dataset was sufficient to provide robust training for the models included. (3) Developability parameters with incomplete values (to avoid training the models on NA values) and one single parameter from each doublet (to avoid training the models on highly-associated parameters) were excluded from the predictability analysis as detailed in Supp Table 2.
Replication	Computational replications and iterations are detailed within the manuscript where relevant.
Randomization	Antibodies were mainly allocated to isotype (IgD, IgM, IgG, IgA, IgE, IgK, IgL), species (human or mouse) and chain type (heavy or light) based on antibody-speciific metadata accompanied with each antibody sequence, or produced during data preprocessing. Developability parameters were classified into sequence-based or structure based parameters depending on the output used to compute them.
Blinding	As the analyses mentioned in this study were conducted in a data driven manner, researchers were not blinded by the way of knowing which developability parameters were sequence-based or structure-based, and to which species, isotype and chain type each antibody sequence belonged to.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A