Absolut! synthetic antibody-antigen binding database.

This dataset contains computed structures from human and murine CDRH3 sequences to 159 PDB-derived lattice antigens.

The dataset has been generated using the Absolut! framework, available at:
https://github.com/csi-greifflab/Absolut

Please refer to / cite this manuscript for general explanations
Robert et al. 2021, A billion synthetic 3D-antibody-antigen complexes enable unconstrained machine-learning formalized investigation of antibody specificity prediction. [will be available on biorxiv in ~July 2021]

Structure of the dataset: (please refer to the github repository for file formats explanation)

****** Absolut! database: The 1 billion antibody-antigen structures (6.9e6 murine CDRH3s >= 11AAs) ******

RawBindingsMurine/

UniqueCDR3s.txt => list of used murine CDRH3 sequences [Greiff 2018 Cell Reports, PMID: 28514665]

XXXX_X.zip => for antigen XXXX_X (example 1ADQ_A = antigen generated from PDB 1ADQ, chain A),
the energetically optimal binding structure of each 11-mer from each CDRH3 to this antigen.
the CDRH3 sequences are separated into multiple text files that can be concatenated.
[careful, the column header "Best?" is written as "Best" only in some files
(unnoticed change of format happened)]

****** Filtered sequences from the database according to affinity thresholds ******

(this allows to work only on sequences with high affinity => smaller files)

RawBindingsPerClassMurine/

XXXX_XAnalyses/ => filtered sequences to antigen XXXX_X (examples below for 1ADQ_A)

*** CDRH3 and 11-mer(slice) -based ***

1ADQ_A_500kNonMascotte.txt => 500k sequences randomly sampled from non-binders (top 99% energies)

*** CDRH3-based top sequences (only keeps the best 11-mer from each CDRH3 to describe its binding) ***

1ADQ_A_superHeroes,txt => bottom 0.01% energies

1ADQ_A_Heroes.txt => bottom 0.1% energies

1ADQ_A_Mascotte.txt => bottom 1% energies (= top 1% high affinity)

1ADQ_A_Looser.txt => bottom 5% energies

The files with "Exclusive" mean excluding the higher affinity class

example: 1ADQ_A_MascotteExclusive.txt = bottom 0.1% to 1% (excludes heroes and super heroes)

*** 11-mer based top sequences (threshold defined from the bottom CDRH3 sequences, but keep multiple low energies 11-mers from the same CDRH3 if they satisfy the threshold (not only the best per CDRH3) - see biorxiv paper above ***
Files havs identical names but including "Slices", example:
1ADQ_A_MascotteSlices,txt

****** Pre-processed datasets used in the Robert 2021 Biorxiv mentioned above: ******

Datasets1/ [binary classification per antigen]
- 11mer-based/ => (used in the manuscript)
  - 1ADQ_A_Task1_SlicesBalancedData.txt => contains all the bottom 1% sequences (Mascotte) to the antigen + the same amount of non-binders sampled to originate from same CDRH3 length distribution
- CDRH3-based/ => also provided but not used (same generation procedure)

Datasets2/ [binary classification per antigen]
- 11mer-based/ => (used in the manuscript, harder: separating bottom 1% to bottom 1% to 5%)
  - 1ADQ_A_Task1_SlicesBalancedData.txt => contains all the bottom 1% sequences (Mascotte) to the antigen + the same amount of bottom 1% to 5%, sampled to originate from same CDRH3 length distribution
- CDRH3-based/ => also provided but not used (same generation procedure)

Datasets3/ [multi-class classification for subsets of N antigens]
- nonRedundant_11mer-based/ => using (142) antigens generated from different-proteins in total
  - Treated142.txt => Binding profile of each bottom 1% CDRH3 to each antigen (column) => 142 columns
  - ListAntigens142.txt => name/meaning of columns of Treated142.txt in this order
  - Task2Annotated_142_nonredundant.zip => annotation of each sequence with the status "non-binder (top 99%)" or which antigens it bind if binder (bottom 1%). this was used to generate Treated142.txt, provided for information.
- redundant_11mer-based/ => using (159) all antigens even if they were generated from the same protein (different PDB)
- Task2V6Apr.py => this script takes TreatedXXX.txt and generated the multi-class dataset for N antigens (it just selects randomly N columns (antigens)) and keeps monospecific sequences within these selected antigens. [it also does the ML classification]

Datasets4_Paratope_Epitope/ (list of all paratope-epitope pairs from binders (bottom 1%) according to different encodings.

perEncoding/ (used in the biorxiv manuscript - only unique paratope-epitope pairs)

example:

Task4_A_EpiSeq_ParaSeq.txt => sequence encoding, degree-explicit

Task4_A_EpiSeq_ParaSeqD2.txt => sequence encoding, degree-explicit, filtered with only residues degree 2 or more

Task4_E_EpiSeq_ParaSeq_NoDeg => sequence encoding, degree-free

Task4_E_EpiSeq_ParaSeq_NoDegD2 => sequence encoding, degree-free, filtered with only residues degree 2 or more

... with Motif, Aggregate and Chemical encodings.

Task4_J_ABDB_Motif_StartX.txt => gapped-motifs encodings (ex: X12XXX1 means gaps of size 12 and 1)

perAntigen/

also provided for each antigen separately

allDegreeExplicitFeatures/

the list of all encodings for each CDRH3 (or each 11-mer - filtered degree 2 residues or not)

so this would allow to use different type of encodings for paratope and epitope, and these files can be used to regenerate the files in perEncoding/ or perAntigen/

DatasetsAdditional/

Variations of Datasets1 and 3 with different affinity thresholds:

Hard means comparing Mascotte (1% bottom) versus Loosers exclusive (1% to 5% bottom)

Harder means comparing Heroes (0.1% bottom) versus Mascotte exclusive (0.1% to 1% bottom)

****** Pre-computed list of all possible binding structures to each antigen of the database ******

(these files are needed to calculate binding structure of a new CDRH3 sequence to those antigens, by the Absolut! software.
They can automatically be re-generated by Absolut! but it takes 1 to 5 days on one core. Easier to download from here.)

Structures/

example: 1d6c68fe18082e4c9eaafa0ec9ae7543-10-11-fa152885be4f2bdfa1c07997182f3093Structures.txt

****** Comparison with (1e6) human-derived CDRH3 sequences to 23 antigens ******

RawBindingsHuman/

XXX_X/

XXXX_XFinalBindings_Process_1_Of_1.txt => structurally-annotated binding of ~1 million human CDRH3 sequences to antigen XXXX_X [Dewitt Plos One 2016, PMID: 27513338]
=> use the CDR3 column to get the list of used sequences