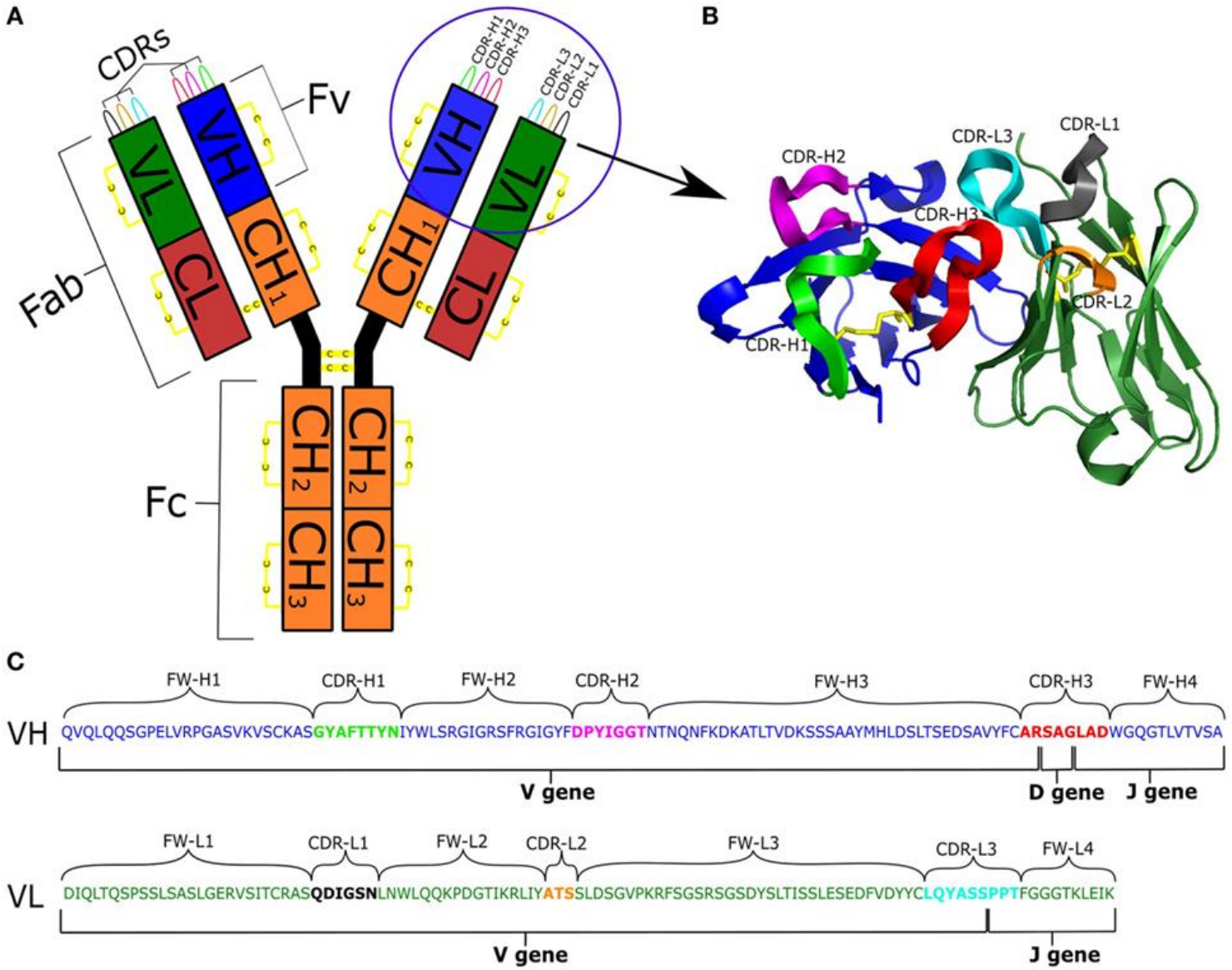
Using structural information to aid in-silico therapeutic design in the era of immunoglobulin repertoire sequencing

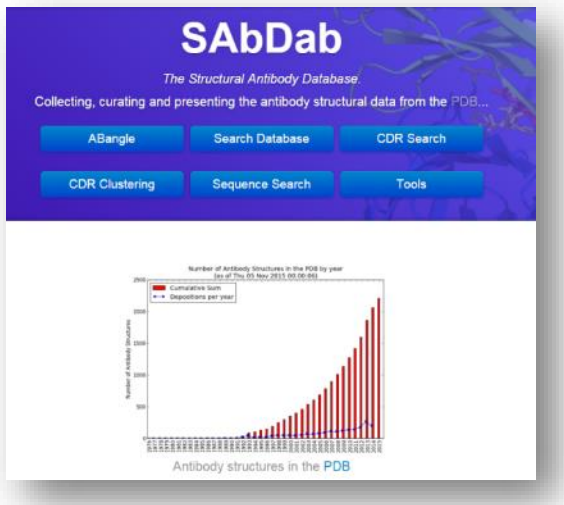
Charlotte Deane
Department of Statistics
Oxford University

Antibody structure



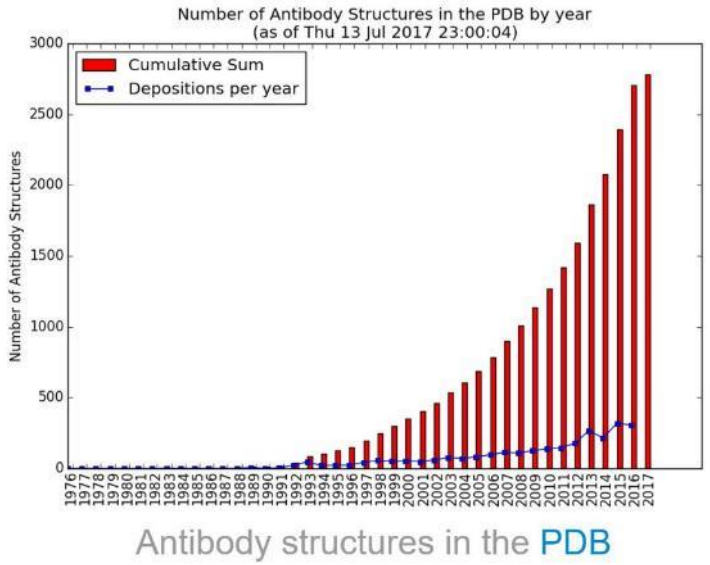
Structural Antibody Database

Public data repository



http://opig.stats.ox.ac.uk/webapps/sabdab

Dunbar *et al.* (2014) *NAR*



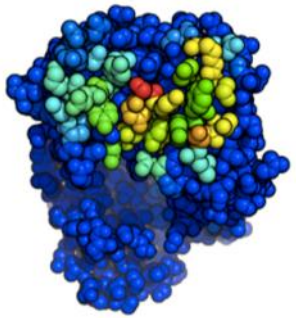
Fully automated updating collection of all publicly available antibody structure data

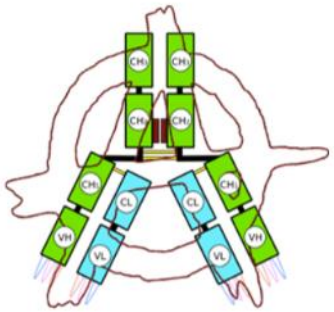
Collect, curate and present structural data consistently.

As of March 22nd 2019

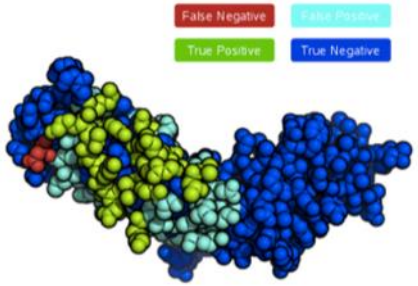
3495 antibody structures

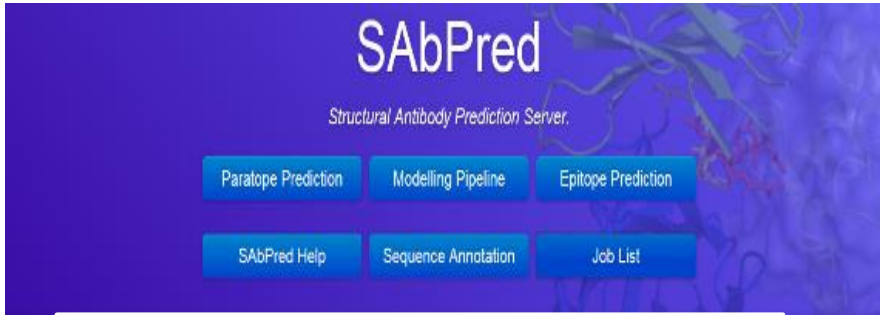
~2398 antibody antigen complexes

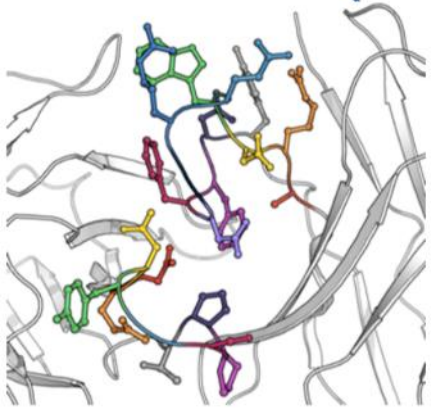




ANARCI: Dunbar & Deane, 2016.

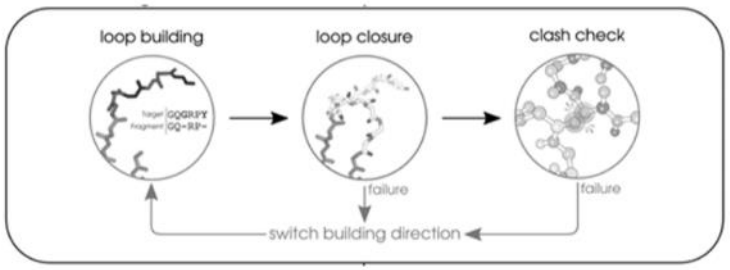






i-Patch, EpiPred:

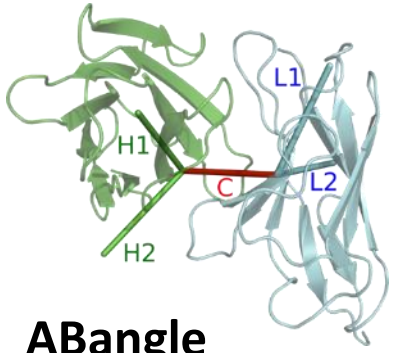
Krawczyk et al., 2013; 2014



Sphinx: Marks *et al.*, 2017.

ABodyBuilder, PEARS:

Leem *et al.*, 2016; 2018

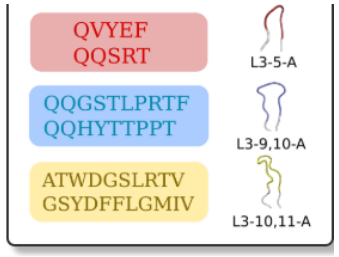


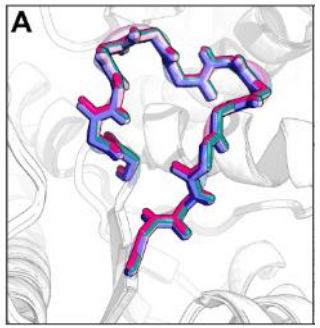
ABangle

Dunbar *et al.* 2013

SCALOP

Wong *et al.*

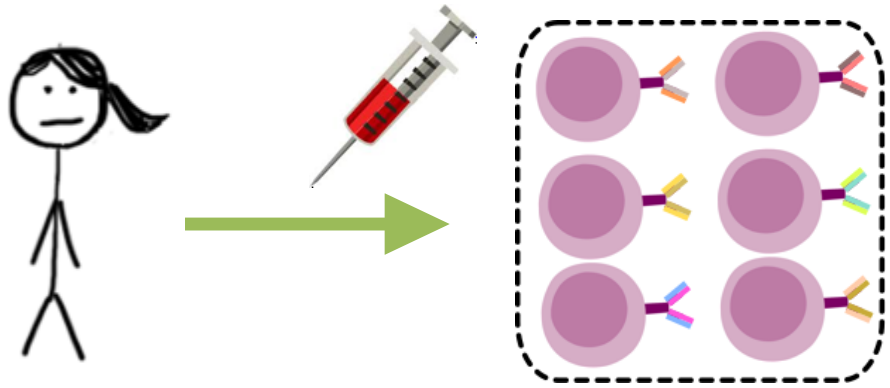




FREAD

Choi *et al.* 2009

Antibody next-generation sequencing (Ig-seq)

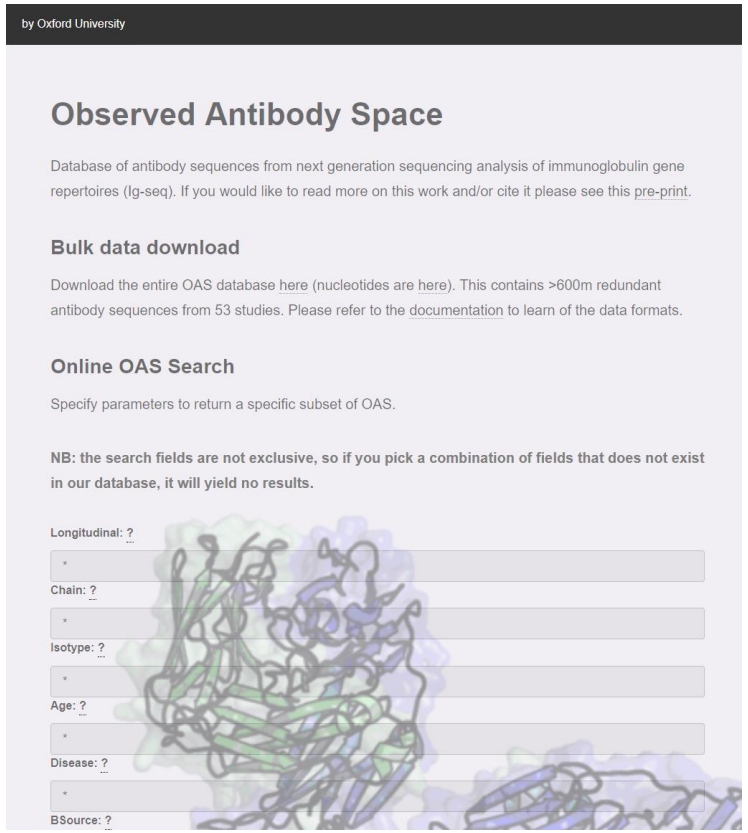


GGTCCCTGAGACTCTCCTGTGCAGCCTCT
GGATTCACCTTTGATGATTATGCCATGCAC
TGGGTCCGCCAAGCTCCAGGGAAGGGCT
GGAGTGGGTCTCAGGTACTAGTTGGAGTA
GTAGTTCCATAGGCTATGTGGACTCTGTGA
AGGGCCGATTCACCATCTCCAGAGACAAC
GCCAAGAACTCCCTGTATCTGCAAATGAAC
AGTCTGAGAGTTGAGGACACGGCCTTATAT
TACTGTGCAAAAGATGTTCTTAGCCGCAGC
TGGCGATATCTTGACCCCTGGGGCCATGGA
ACCCTGGTCACCGTCTCCTCAGCATCCCCG
ACCAGCCCCAAGGTCTTCC

- Snapshots between 10^4 and 10^7 antibody sequences
- Theoretical antibody repertoire >10^{10}
- Naïve human antibody repertoire
- Pre and post immunisation datasets
- Sequences repertoires from different species

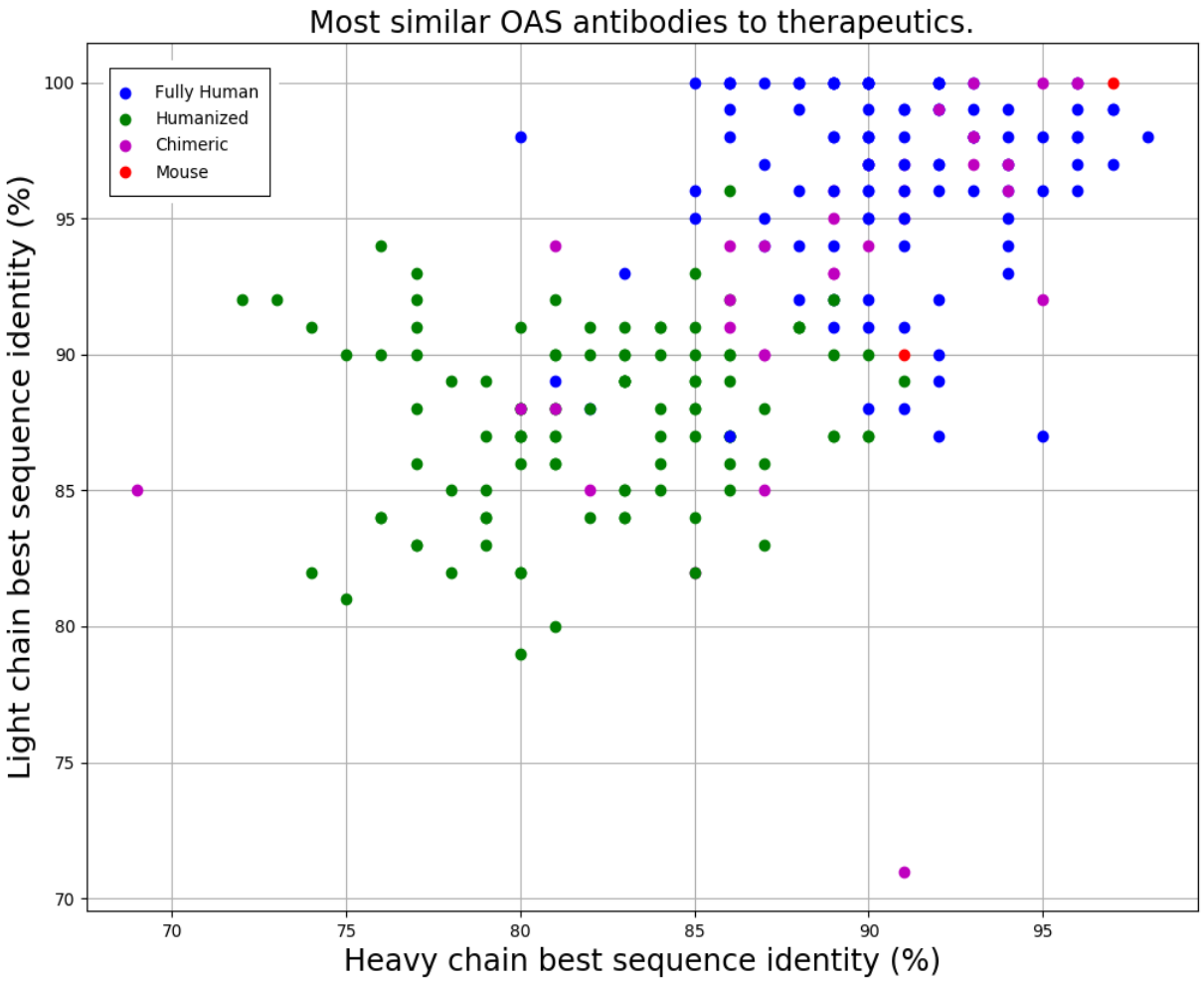
Observed Antibody Space: OAS

- The first organized collection of a large body of antibody NGS outputs
- 60 studies covering ~ 1 billion antibody sequences across diverse immune states, organisms and individuals.
- We have sorted, cleaned, annotated, translated and numbered these sequences and made the data available
- OAS will facilitate data mining of immune repertoires for improved understanding of the immune system and development of better biotherapeutics.
- The data is all made freely available at antibodymap.org



Kovaltsuk *et al*. (2018). *Journal of Immunology*

Looking for Therapeutic Antibodies in OAS



- 242 post phase-1 antibodies
- Unexpected high percentages of sequence overlap with therapeutics
- Many can be found in OAS with sequence identities >95%
- Enfortumab, heavy and light chain have 98% seqID
  - differences H38:N-S, H88:S-Y, L37:G-S, L52:F-L
- 54 have a perfect CDRH3 match
  - 22 of these found in more than one dataset

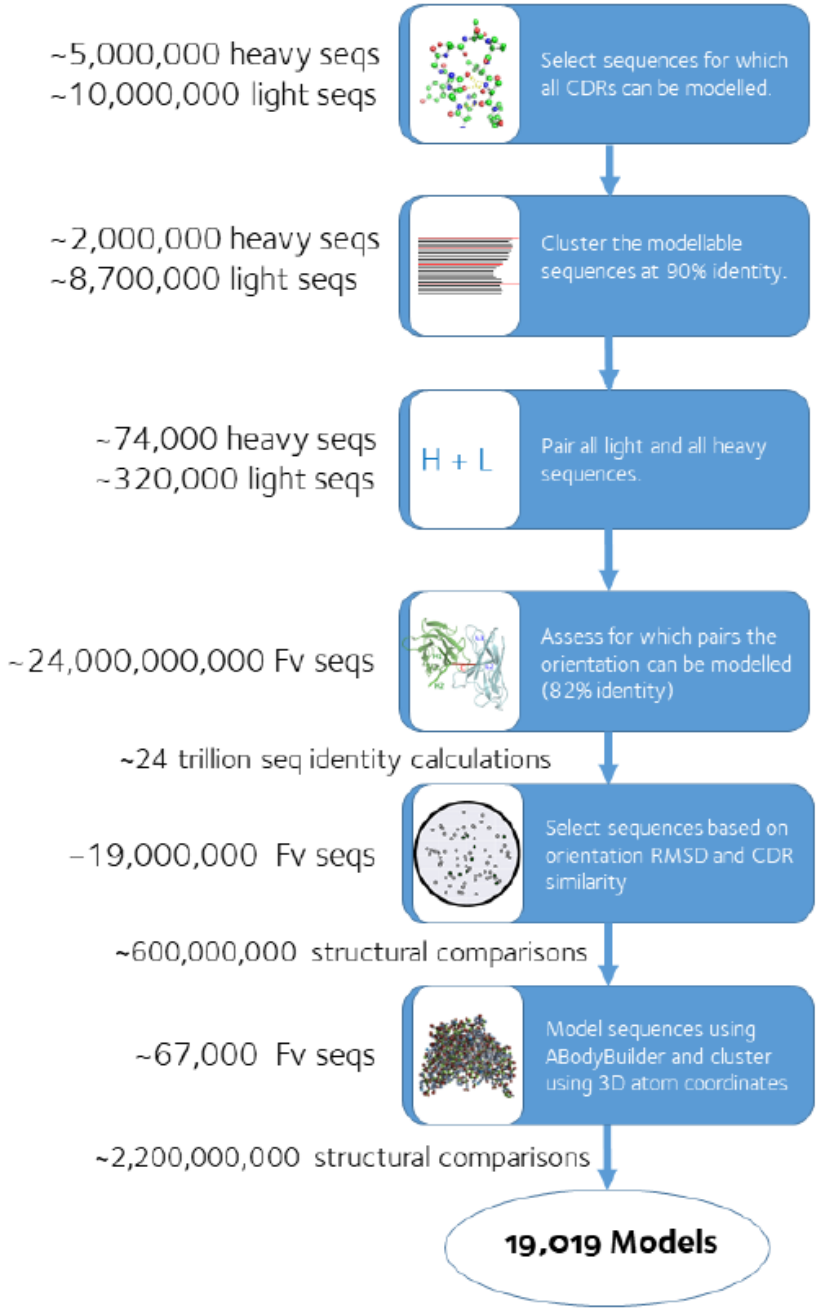
Krawczyk *et al*. (2019). *bioRxiv*

Building a human antibody model library

- Build a database of models that describe the structural variability of human antibody space
  - Input data is unpaired complete VH and VL sequences from the same sample
- Build a pairing and modelling protocol
  - to capture the sequence and structural diversity
  - within the constraints of modelability
  - and computational expense

A Human Antibody Model Library

- Ig-seq data from circa 500 healthy individual
- The sequences are unpaired naïve and memory IgM molecules sourced from peripheral blood, bone marrow and spleen.
- ~13.5 million unique sequences ~5m heavy variable regions and ~8.5m light variable regions.
- Reduces to ~20,000 structurally diverse antibodies that we can model accurately



The Therapeutic Antibody Profiler (TAP)*
Five Computational Developability Guidelines

- Therapeutic antibodies must not only bind to their target but must also be free from 'developability issues' such as poor stability or high levels of aggregation.

- TAP is an *in-silico* antibody design analog of the Lipinski's rule of five for small molecules
  - to guide the selection of antibodies with appropriate biophysical properties

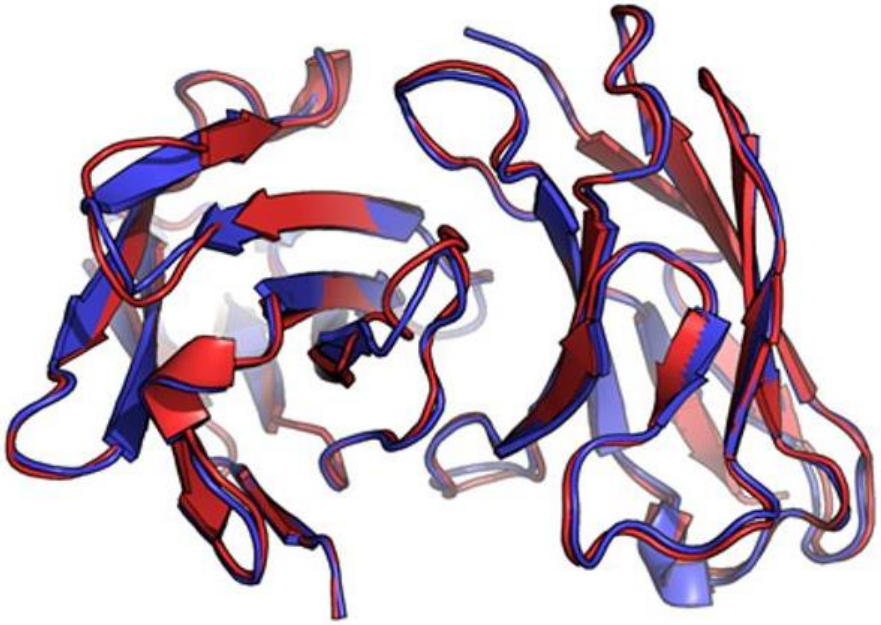
- Derive distributions of metrics for clinical stage therapeutics and assume that these indicate the allowed values of these properties.
  - Calculate these metrics on models so can run against potential therapeutics where crystal structures are unavailable

- These metrics don't have to correlate with a particular experiment that tests for developability rather they indicate that a potential therapeutic has outlying values.

*Raybould et al (2019) PNAS

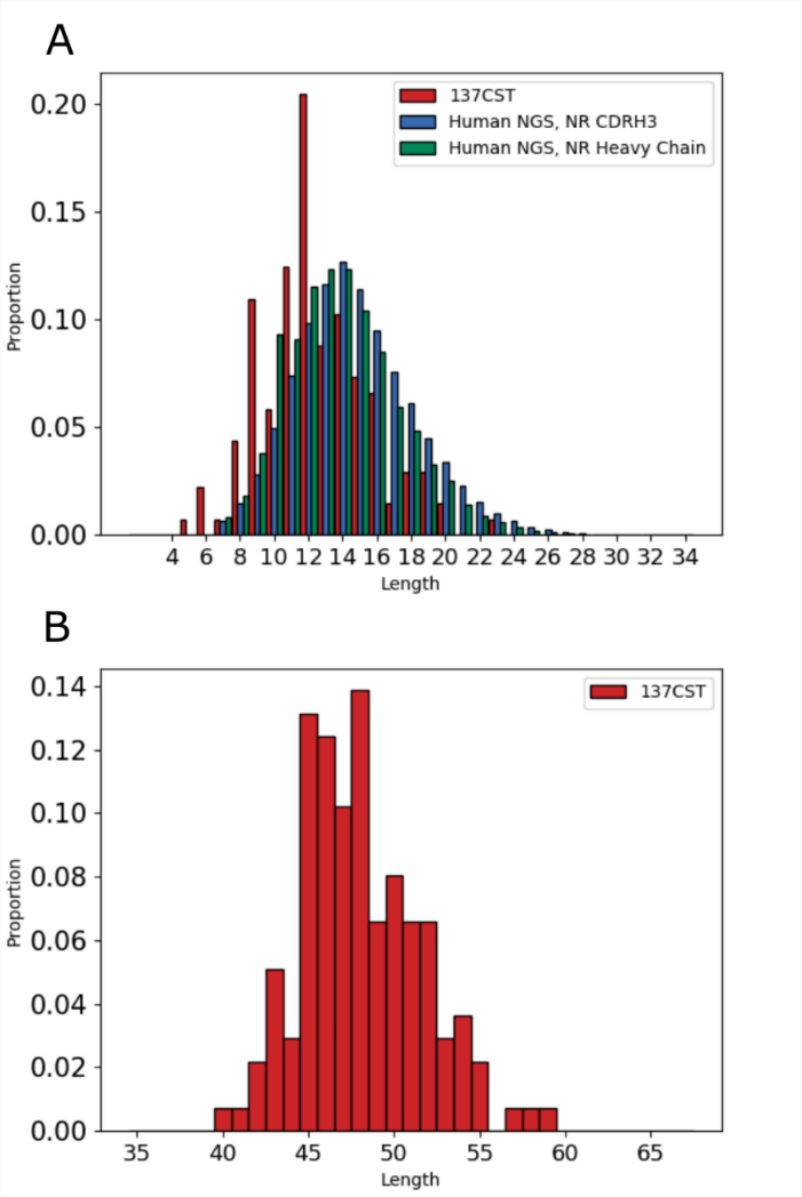
Datasets – structural models

- Models of the variable domain structures of 137 post-Phase I clinical-stage antibody therapeutics (CSTs)*
  - Models are accurate enough for our metrics (tested with the 56 CSTs with known structure)
  - Average RMSD of framework < 1A
  - Less than 4% of residues are wrongly annotated exposed/buried
- For context we use the ~20,000 models of our larger human antibody model library
  - Would any human antibody make a good therapeutic?



*Jain et al (2017) PNAS

CDRH3 length and total CDR length



CDRH3s of our clinical stage therapeutics tend to be shorter than the human Ig-Seq data

So being a human sequence ≠ good therapeutic

Could be for a number of reasons, including **affinity** (reduce entropy loss), or **aggregation**.

Use total CDR length as our metric (across all 6)

To perhaps better capture **binding site shape**
Closely correlates with CDRH3 length ($r = +0.77$, $p < 10^{-28}$)

TAP METRIC 1: TOTAL CDR LENGTH

CDR vicinity Patches of Surface Hydrophobicity (**PSH**)

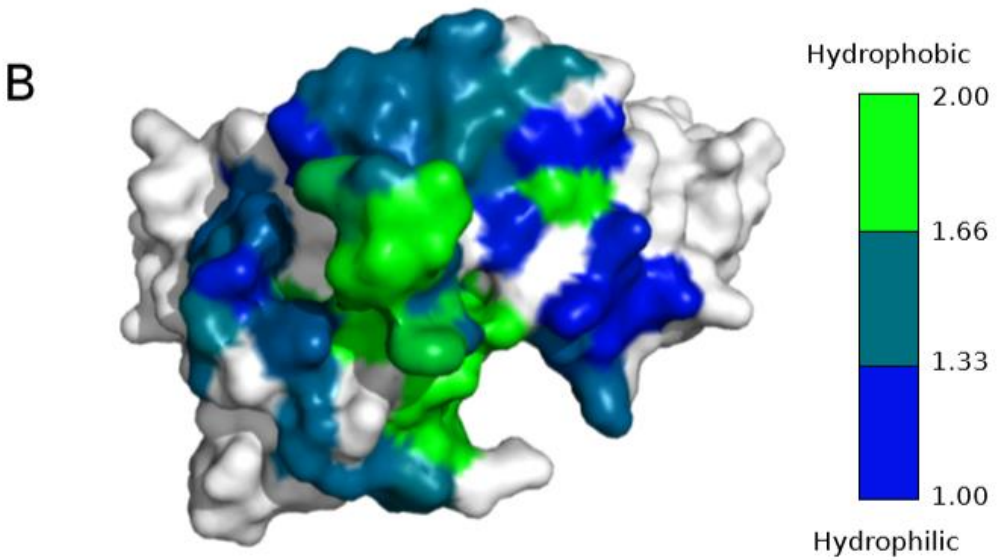
$$\sum_{R_1 R_2} \frac{H(R_1,S)H(R_2,S)}{{r_{12}}^2}$$

R_1,R_2 are different residues
H(R,S) hydrophobicity score
r_{12} – distance between R_1 and R_2

The set of residues considered are the CDR vicinity

All surface-exposed IMGT CDR residues, as well as other surface-exposed residues with a least one heavy atom within a radius of 4Å

TAP METRIC 2: CDR Vicinity PSH



Galiximab, with a high CDR Vicinity PSH value, has a large patch of hydrophobicity in its CDRH3 loop.

CDR vicinity Patches of Surface Hydrophobicity (**PSH**)

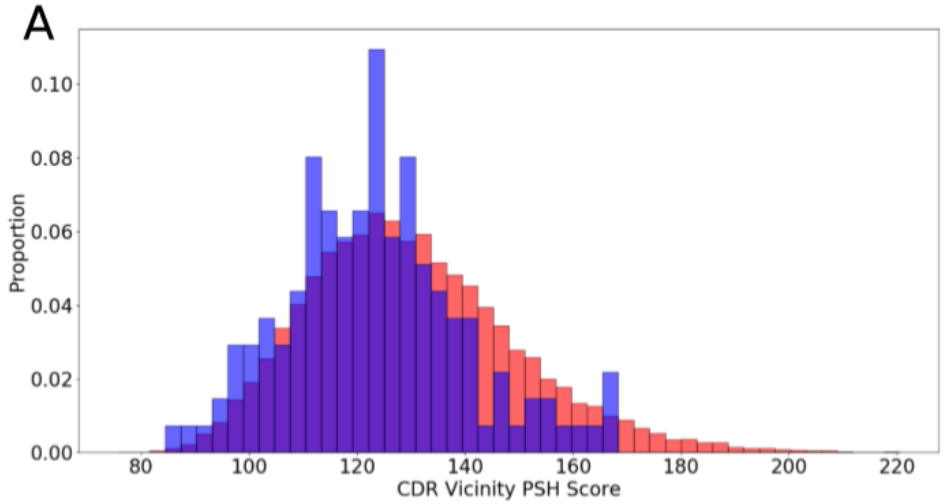
$$\sum_{R_1 R_2} \frac{H(R_1,S)H(R_2,S)}{{r_{12}}^2}$$

R_1,R_2 are different residues
H(R,S) hydrophobicity score
r_{12} – distance between R_1 and R_2

The set of residues considered are the CDR vicinity

All surface-exposed IMGT CDR residues, as well as other surface-exposed residues with a least one heavy atom within a radius of 4Å

TAP METRIC 2: CDR Vicinity PSH

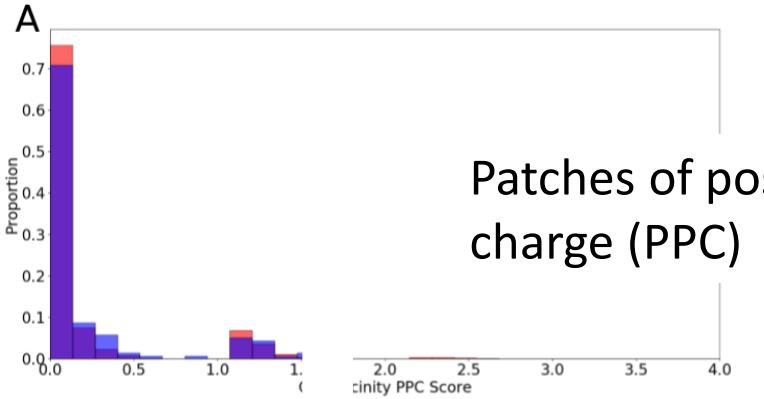


Blue are calculated for the models of the Clinical stage therapeutics

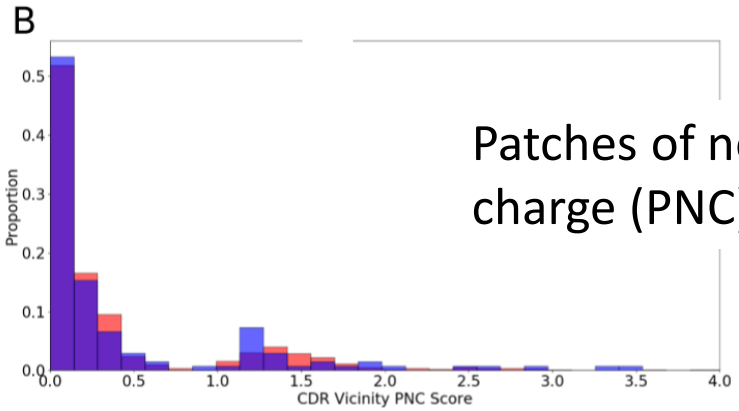
Red are the values for the Human antibody model library

Once again human sequence ≠ good therapeutic

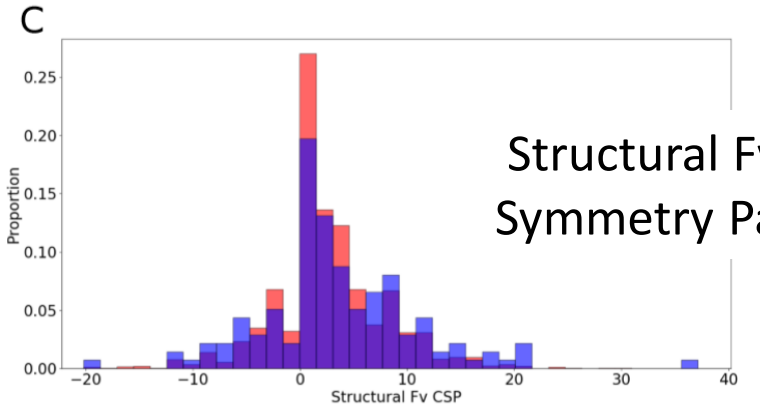
Charge



Patches of positive charge (PPC)



Patches of negative charge (PNC)



Structural Fv Charge Symmetry Parameter (SFvCSP)

TAP METRIC 3: CDR Vicinity PPC

$$\sum_{R_1 R_2} \frac{|Q(R_1)|\,|Q(R_2)|}{{r_{12}}^2}$$

Q(R) charge of residue R

TAP METRIC 4: CDR Vicinity PNC

$$\sum_{R_H} Q_{R_H} \sum_{R_L} Q_{R_L}$$

RH, RL are surface-exposed VH, VL residues respectively.

TAP METRIC 5: SFvCSP

Five selected metrics – amber and red flags

- Total CDR Loop Length

Longer lengths ∝ Aggregation, Lower Affinity, Convex Binding Site

- Patches of Surface Hydrophobicity (PSH) in the CDR Vicinity

Higher values ∝ Aggregation (Self-association), High Viscosity

- Patches of Positive Charge (PPC) in the CDR Vicinity

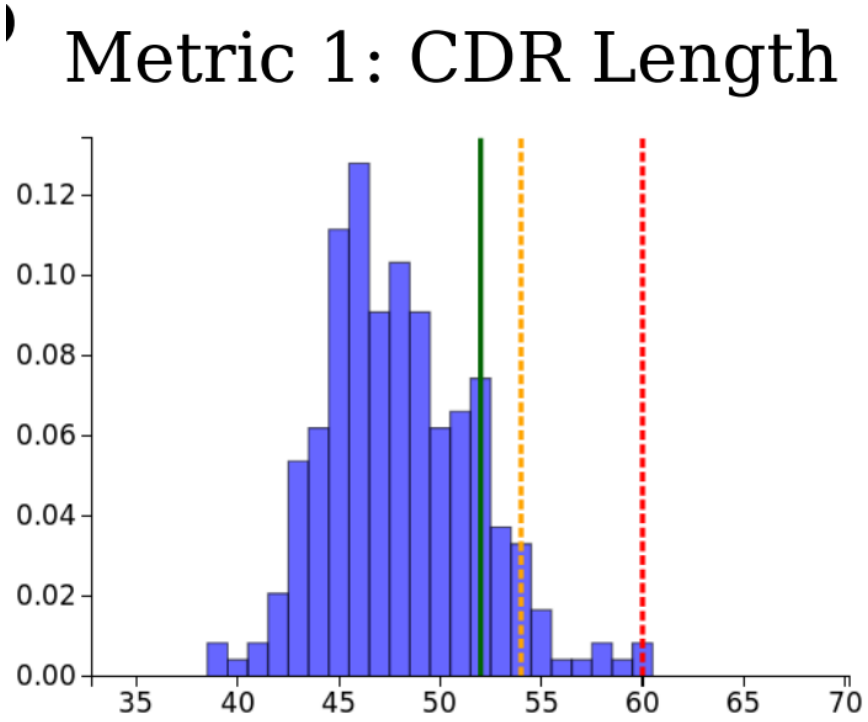
Higher values ∝ Poor Expression/Stability

- Patches of Negative Charge (PNC) in the CDR Vicinity

Higher values ∝ Poor Expression/Stability

- Structural Fv Charge Symmetry Parameter (SFvCSP)

More negative values ∝ Colloidal instability, High Viscosity



Amber within 5% of highest/lowest value seen
Red above any value seen

Validation: Things TAP shouldn't flag

- Tested against 105 extra post-Phase I therapeutics

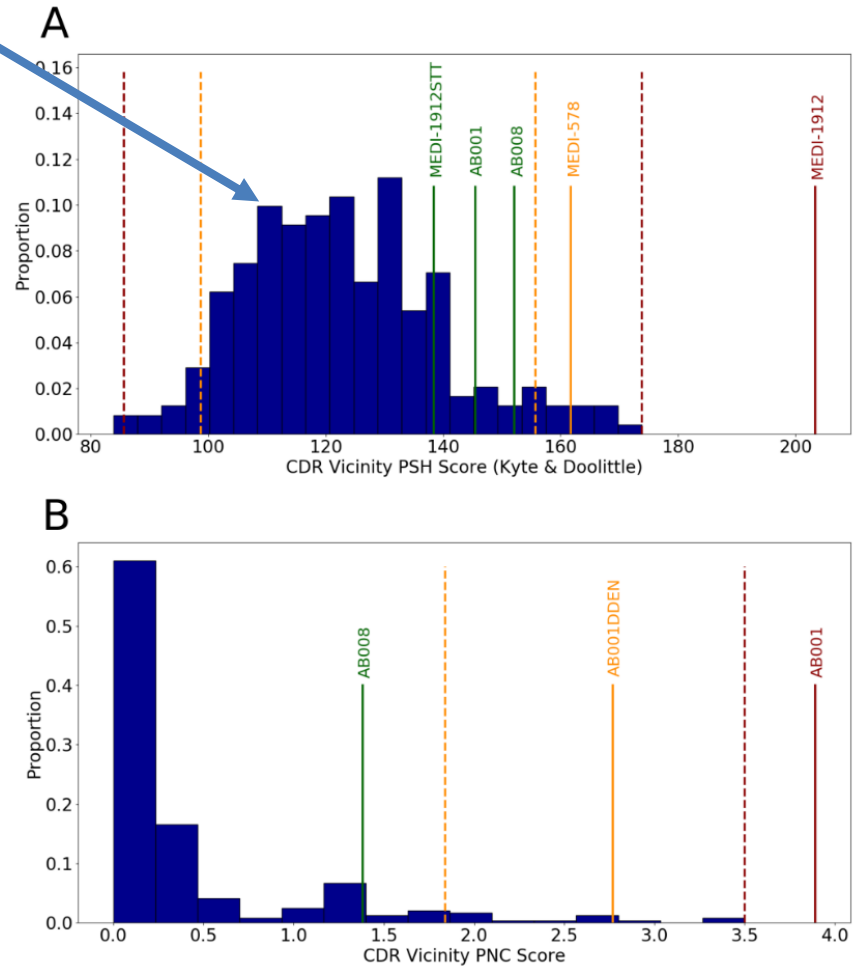
Metric	137 CST Amber Flag Region	Number Amber Flagged	137 CST Red Flag Region	Number Red Flagged
Total CDR Length (L)	$54 < L \leq 59$	6	$L > 59$	2*
PSH, CDR Vicinity (Kyte)	$85.65 \leq PSH < 98.74$	2	$PSH < 85.65$	1
	$155.76 < PSH \leq 171.91$	5	$PSH < 171.91$	1*
PPC, CDR Vicinity	$1.23 \leq PPC < 1.51$	1	(> 1.51)	5*
PNC, CDR Vicinity	$1.90 \leq PNC < 3.50$	4	(> 3.50)	0
SFvCSP	$-39.00 \leq SFvCSP < -18.00$	1	(< -39.00)	1

*Erenumab flagged for each of these properties

- Low red-flagging rate (8 of 105), implies won't pick out genuine therapeutics as having issues very often.

Validation: Things TAP should flag

242CSTs



Anti-NGF antibody (1)
MEDI-578: Minor aggregator;
MEDI-1912: Extreme aggregator;
MEDI-1912STT: No aggregation.

TAP predicts a similar behaviour from just the sequence and its ABodyBuilder model

Anti-IL13 candidate (2)
AB008: Good expression;
AB001: Poor expression (7x lower than AB008);
AB001DDEN: Good expression.

TAP flags AB001 for further consideration.

(1) Dobson et al (2016) Sci Reports
(2) Popovic et al (2017) Protein Eng Des Selec

Splitting by Species Origin

Table S8. 242 CST TAP values split by species origin.

TAP Metric	101 Human ($\mu \pm \sigma$)	108 Humanized ($\mu \pm \sigma$)	30 Chimeric ($\mu \pm \sigma$)	3 Mouse ($\mu \pm \sigma$)
Total CDR Length	48.68 ± 4.09	47.80 ± 3.42	46.77 ± 3.55	46.33 ± 1.25
PSH	127.76 ± 18.56	120.90 ± 14.20	115.73 ± 15.58	117.26 ± 9.44
PPC	0.29 ± 0.58	0.20 ± 0.36	0.26 ± 0.55	0.05 ± 0.06
PNC	0.34 ± 0.56	0.50 ± 0.75	0.30 ± 0.63	0.50 ± 0.50
SFvCSP	4.06 ± 7.44	3.13 ± 7.80	3.29 ± 5.99	7.58 ± 6.75

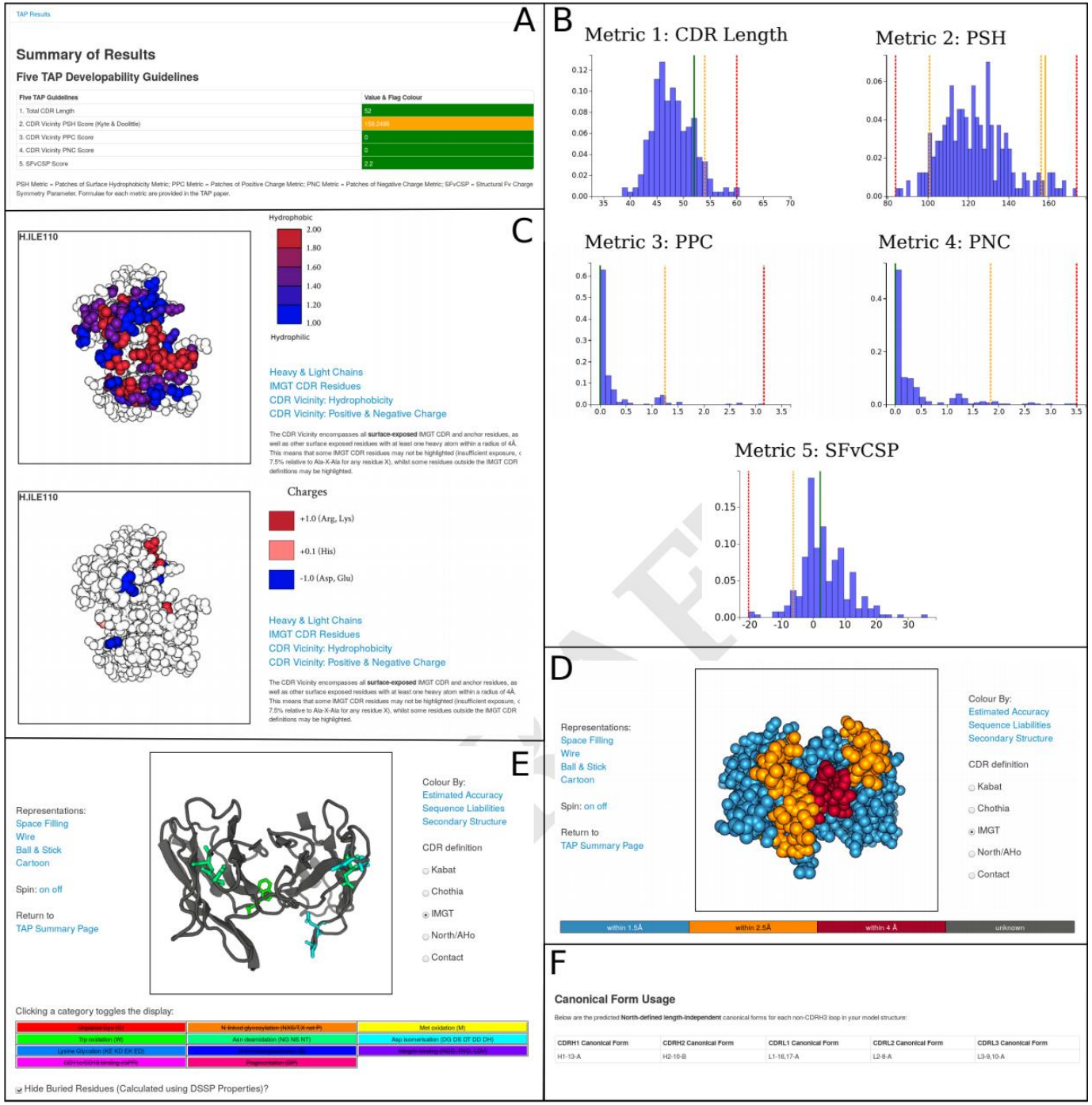
- Appears that the more human mAbs have larger patches of hydrophobicity than mouse mAbs
- We also split by clinical progression (P2, P3, Approved) and drug campaign status (active/discontinued) but found no significant differences in TAP metric values.

TAP website

Distributions will become more accurate as model quality improves (more up-to-date version of SAbDab) or as more therapeutics come to Phase-II of clinical trials.

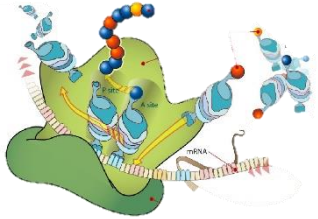
http://opig.stats.ox.ac.uk/resources

Raybould et al (2019) PNAS



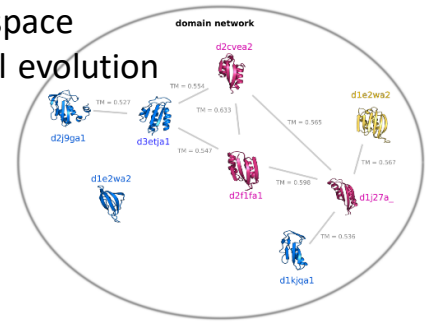
The Oxford Protein Informatics Group

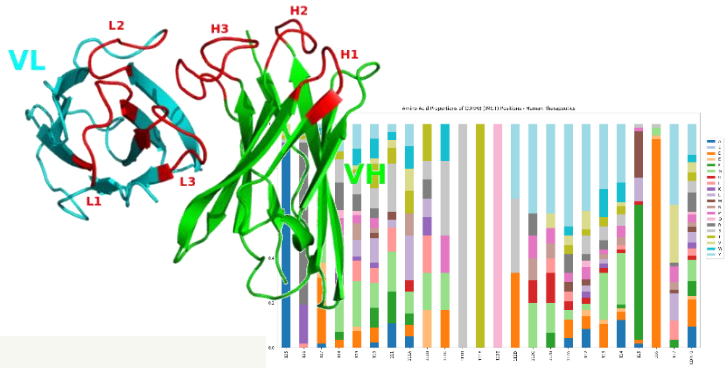
“more than just antibodies”



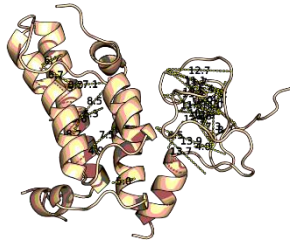
Biologically inspired protein structure prediction

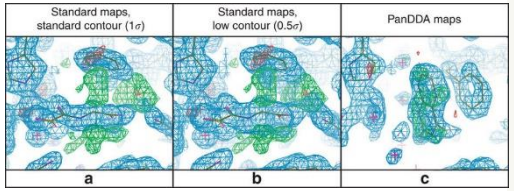
Fold space
& structural evolution



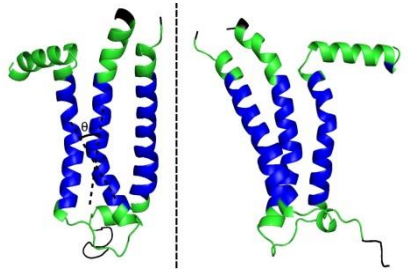


Immunuo Informatics

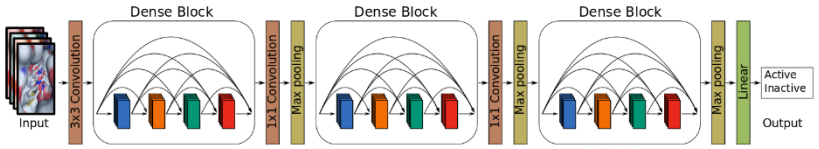




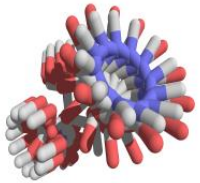
Developing software for Crystallography



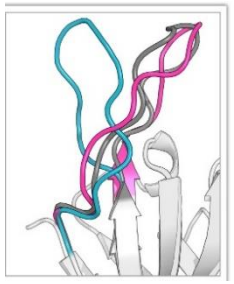
Membrane proteins

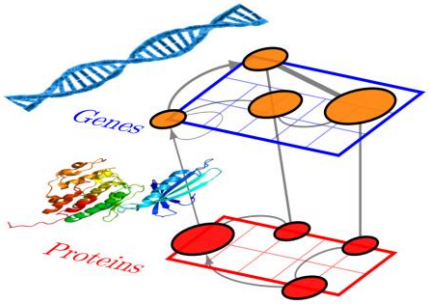


Small molecule drug discovery



Loop structure
Conformations
/Prediction

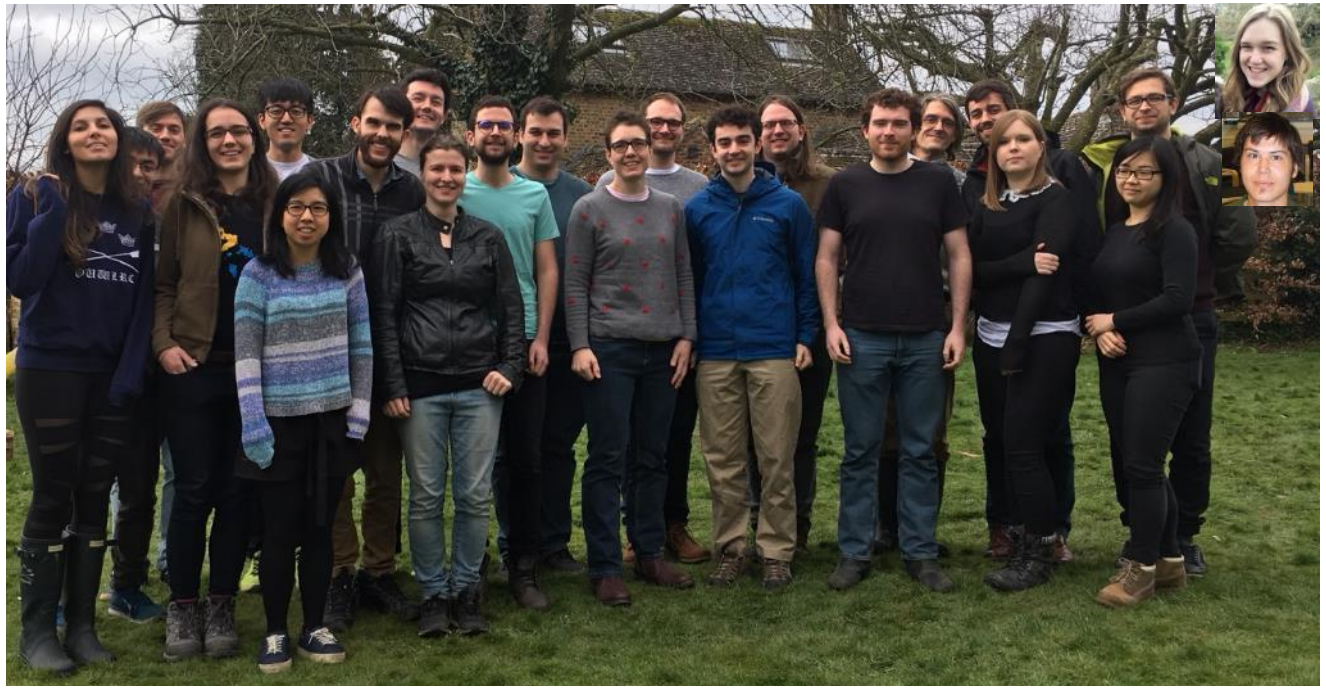




Networks

UNIVERSITY OF OXFORD

ACKNOWLEDGEMENTS



MRC Medical Research Council

kymab

BenevolentAI

NOVARTIS
NOVARTIS INSTITUTES FOR BIOMEDICAL RESEARCH

Exscientia
DRIVEN BY KNOWLEDGE

MedImmune
A member of the AstraZeneca Group

Roche

diamond

ucb

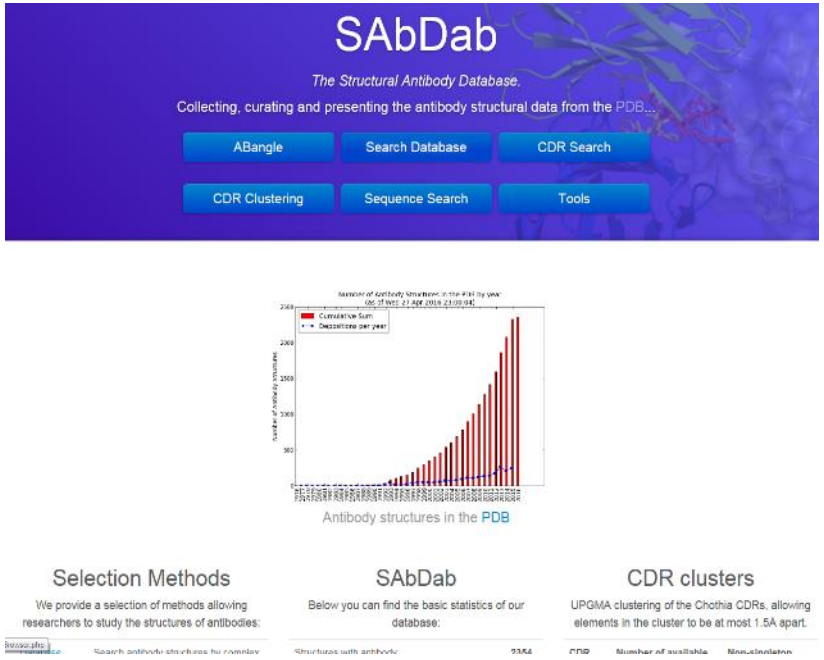
gsk GlaxoSmithKline

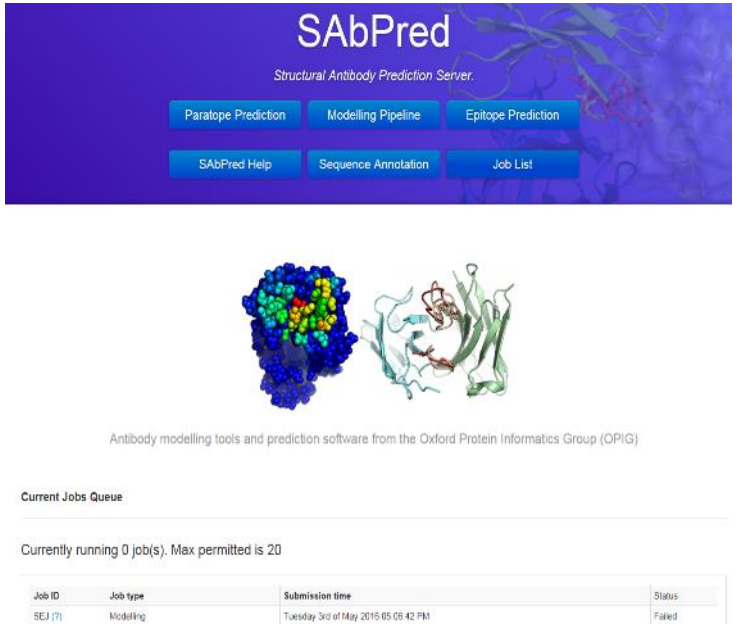
e-Therapeutics plc

EPSRC
Pioneering research and skills

BBSRC
bioscience for the future

SAbDab & SAbPred – a computational antibody design platform





All our software is open and free to use

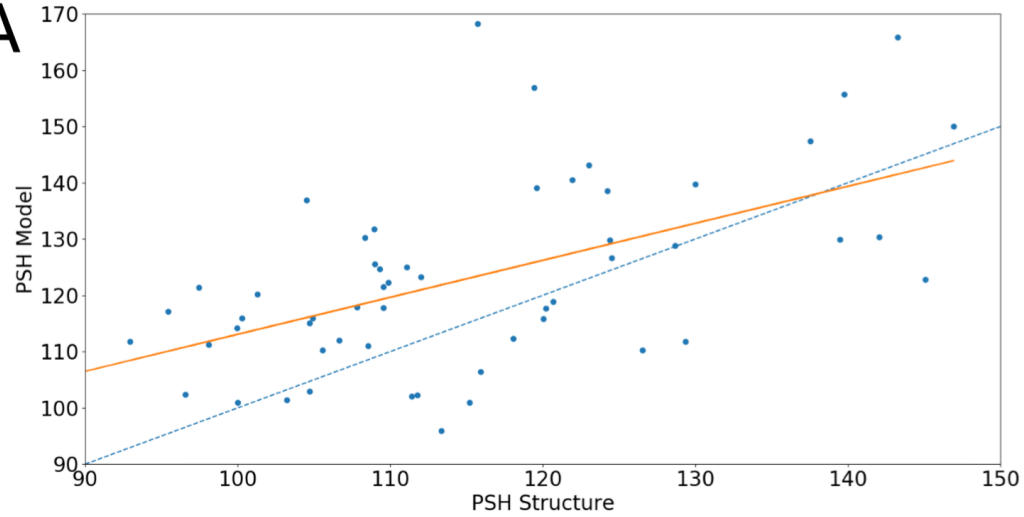
We offer consulting to create custom install specific for your infrastructure and/or training on the platform.

Newly available a VirtualBox virtual machine encompassing SAbDab/SAbPred.

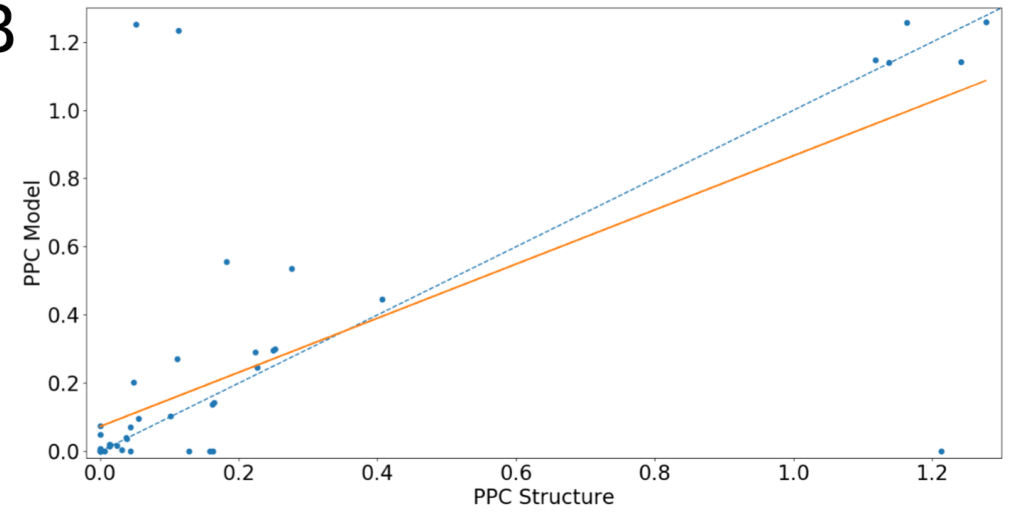
More information contact deane@stats.ox.ac.uk

Values for models vs real structures

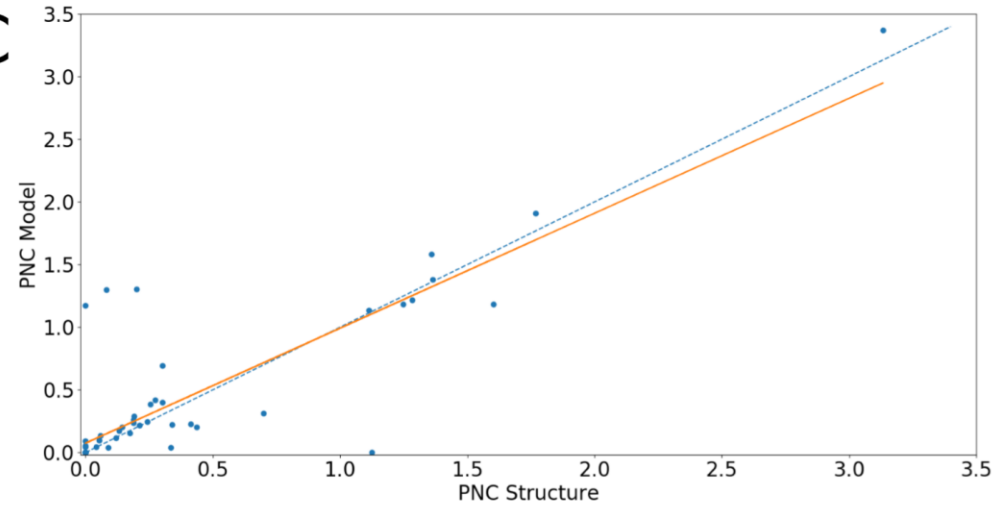
A



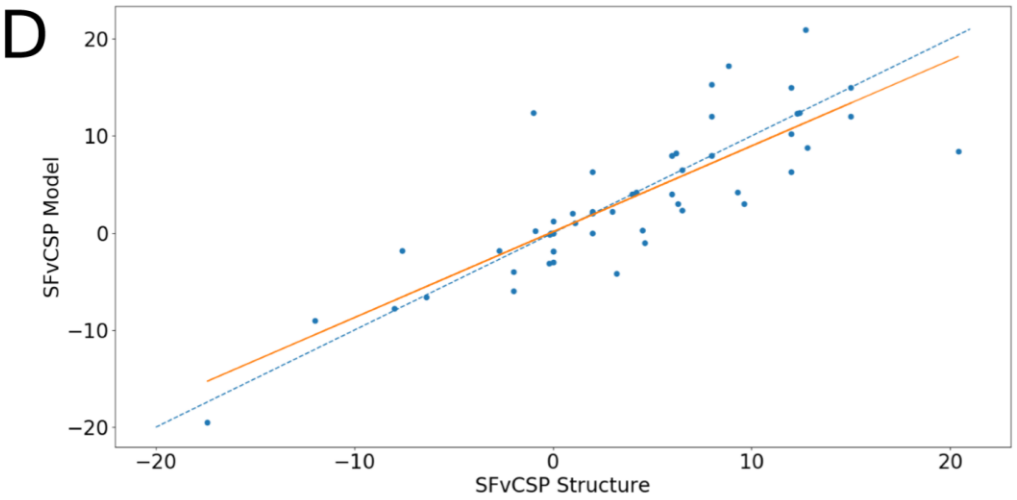
B



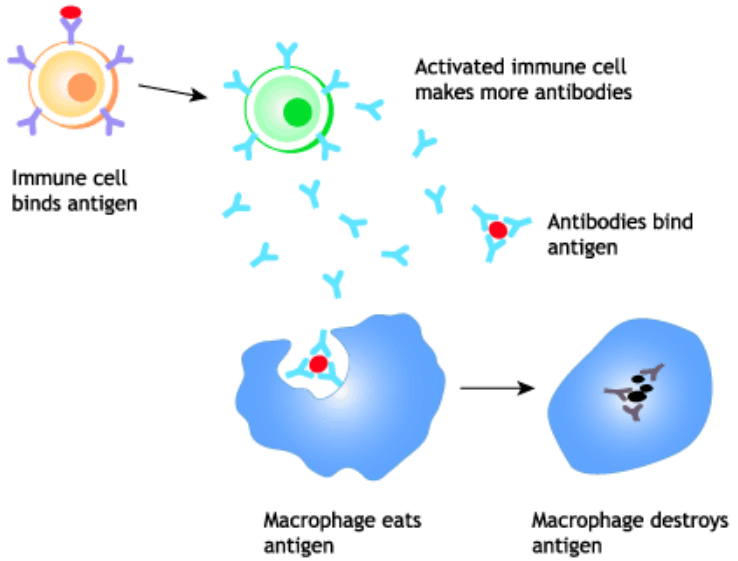
C



D



Antibodies

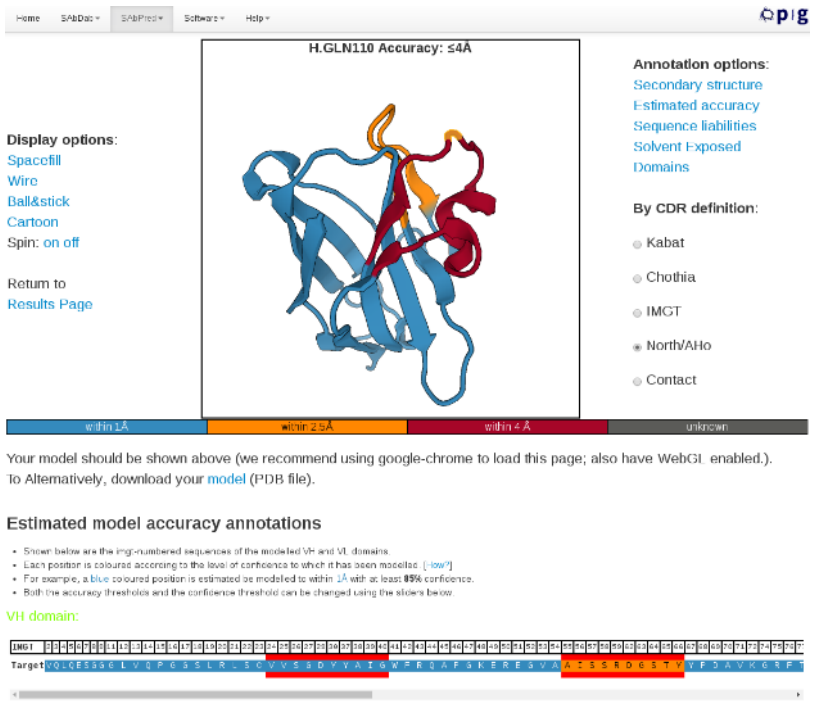


- It has been estimated that a typical human is capable of producing more than 10^{10} different antibodies, each capable of binding a distinct epitope
- Recognise and bind to potentially harmful molecules (antigens)
- Either inhibit the antigen themselves or recruit other parts of the immune system to deal with them

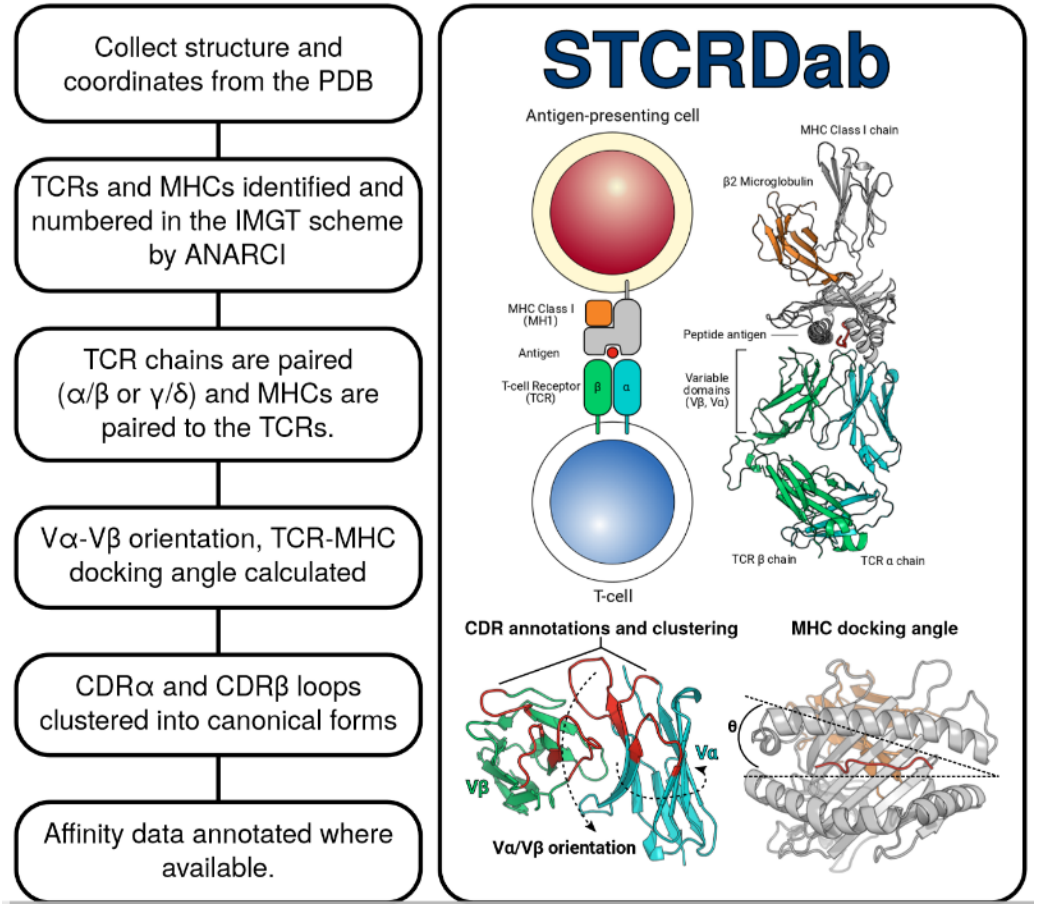
- Target specifically and with high affinity
- Can be raised against almost any antigen
- Effective treatment for immune disorders, cancers and arthritis
- Currently >60 approved antibody “drugs”
- Hundreds in late stage development

Beyond antibodies

Nanobodys



TCRS



Leem at al NAR (2017)