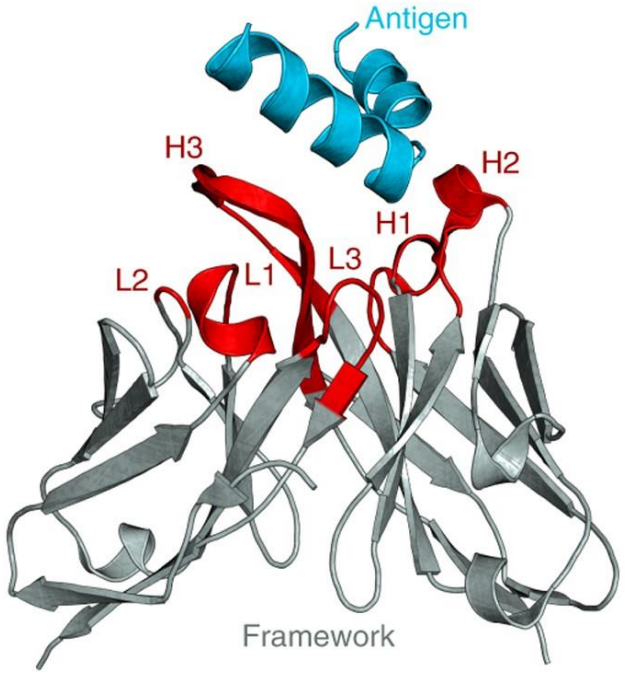
Careers in AI Drug Development



About Me

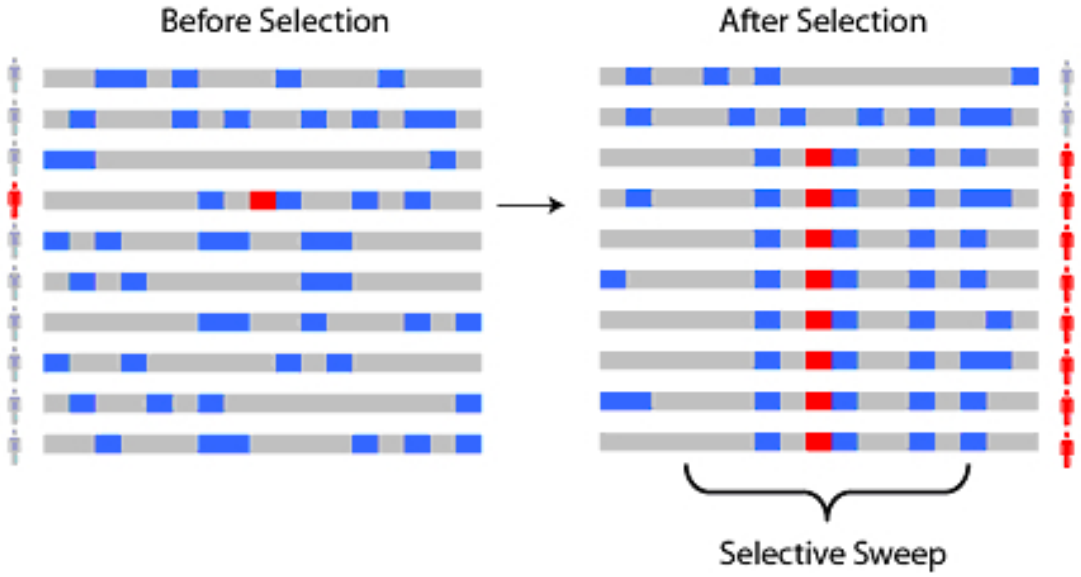
The Royal Veterinary Colleg

- BSc Bioveterinary Science
- MRes Functional Genomics

UCL

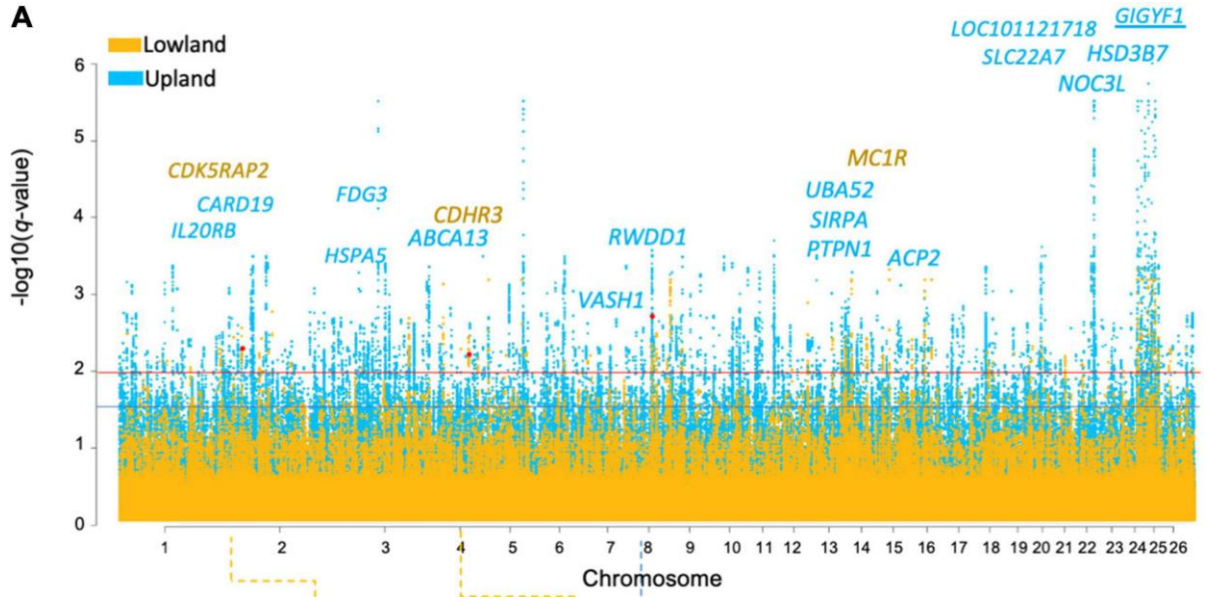
- PhD Antibody developability

What Got Me Started?

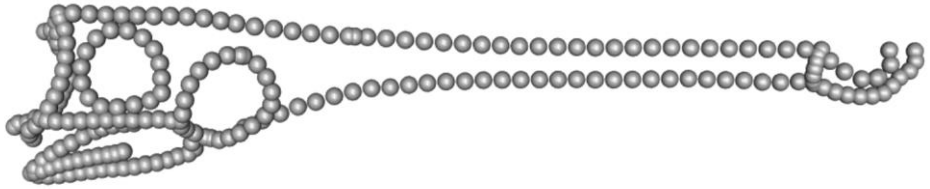


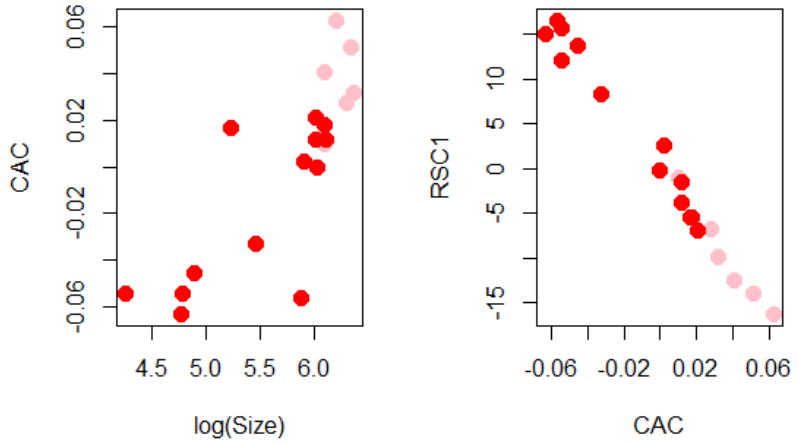
Upland

Lowland

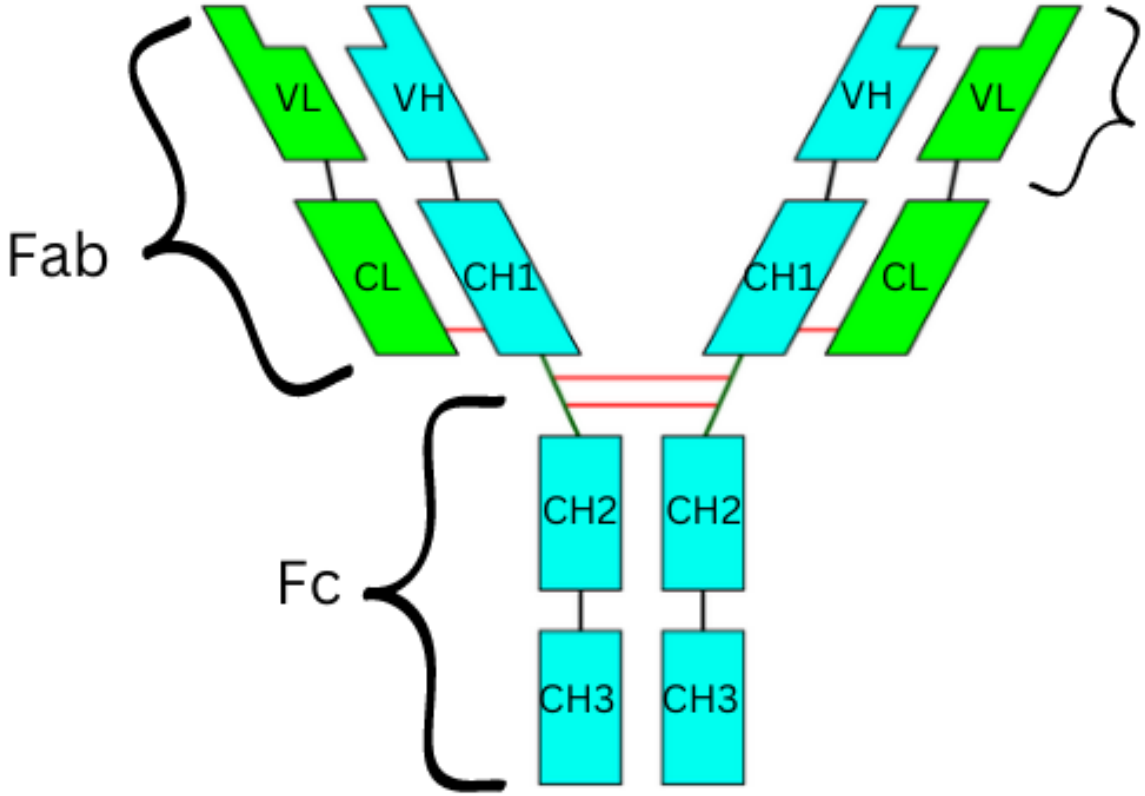


What Got Me Started?

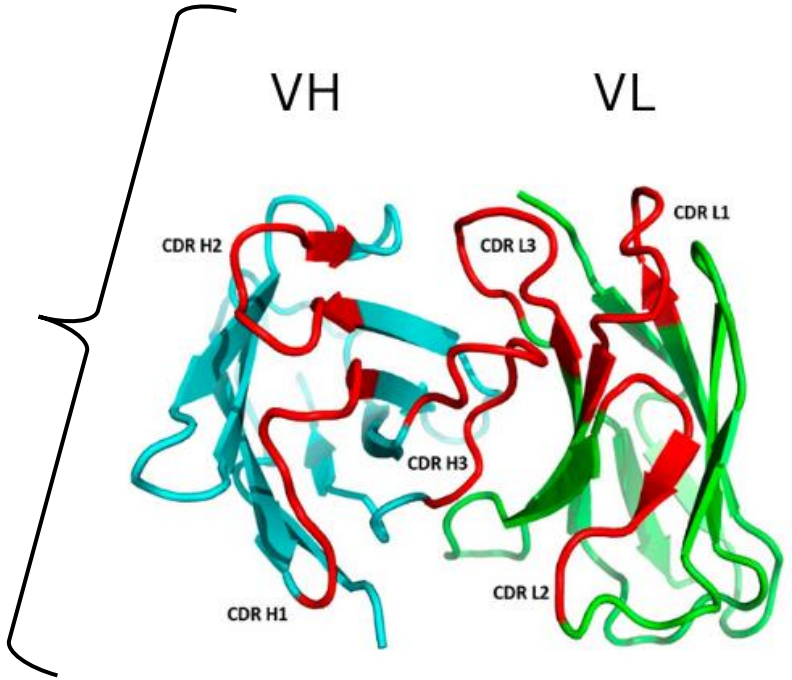




PhD – Monoclonal Antibodies (mAbs)



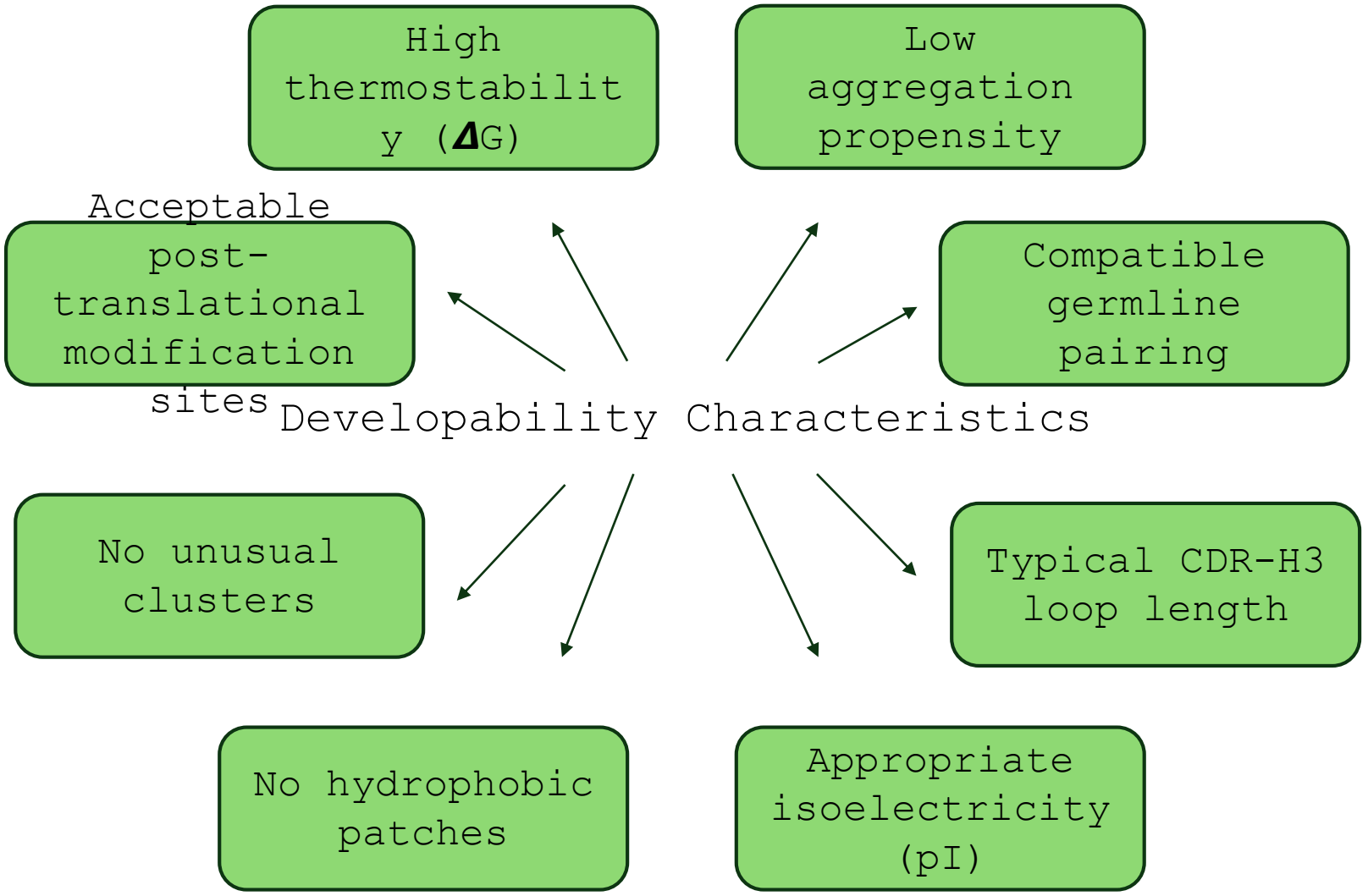
IgG

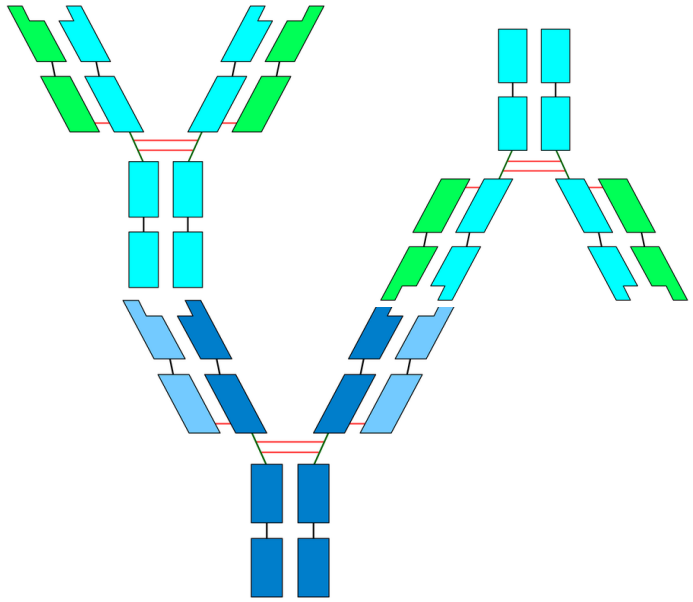


>AAACCTGAGACCGGAT-6_VH
QVQLVQSGAEVKKPGASVKVSCKASGYTFTVFYIFWVRQA
PGQGPEWMGWINPNSGGTSYAQNFQGRVTMTRDTSVSTAY
MELSRLTSDDTAVYFCARGRRGLITEFDYWGQGTLVTVSS

>AAACCTGAGACCGGAT-6_VL
DIVMTQSPDSLAVSLGERATINCKSSQSVLDSSNNKNYLA
WYQQKPGQPPNLLIYWASTREYGVPDRFSGSGSGTDFTLT
ISSLQAEDVAVYYCQQYSSTPFTFGQGTKLEIK

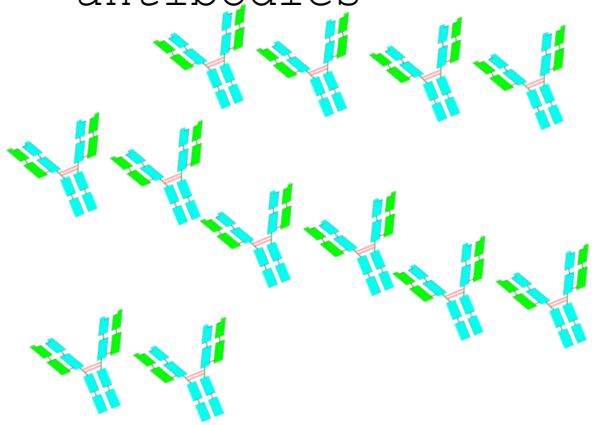
mAb Developability

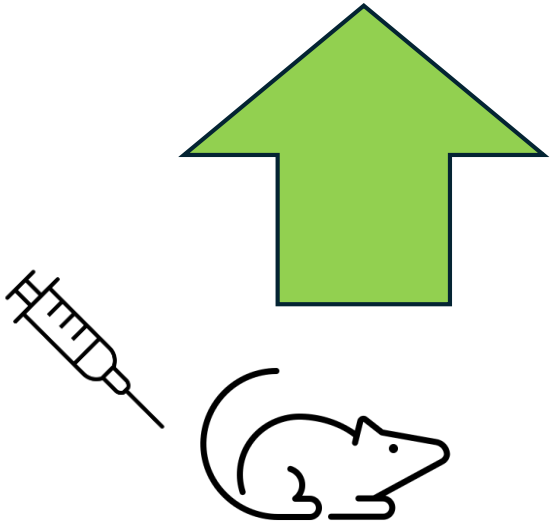




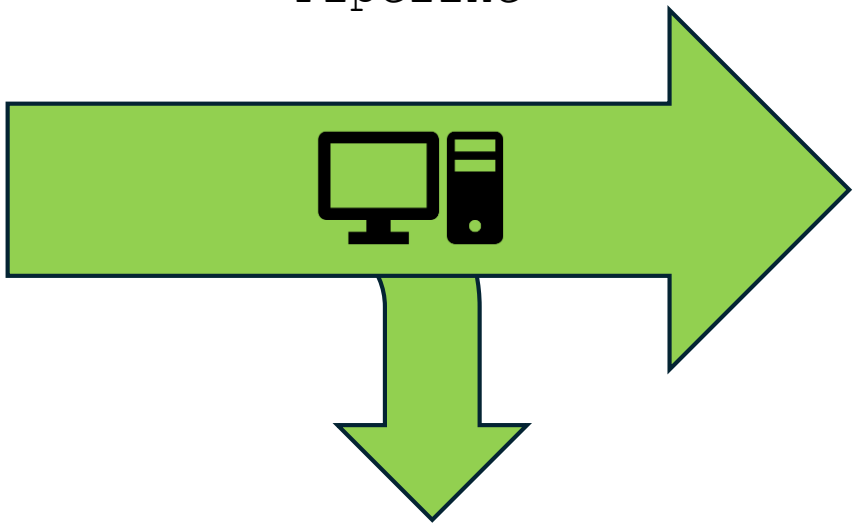
Hypothesis

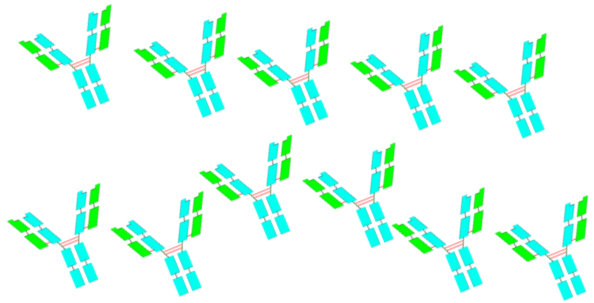
Repertoire antibodies



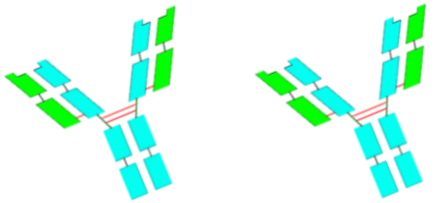


Triaging Pipeline

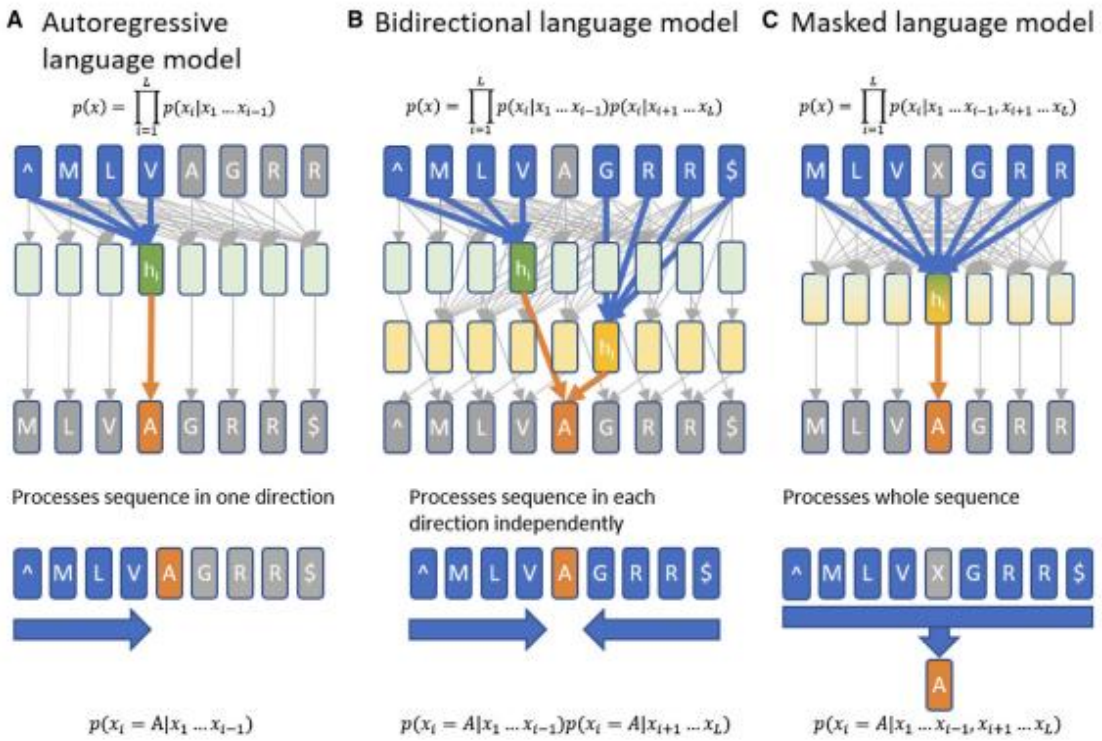




Antibodies with developable features that could be clinical



Antibody Language Models



Language Model	Encodings	Padding Character	Reference
AntiBERTy	n x 512	"_"	Ruffulo *et al.* (2023) *Nature*
AbLang	n x 768	"*"	Olsen *et al.* (2022) *Bioinformatics Advances*
Sapiens	n x 600	"*"	Prihoda *et al.* (2022) *mAbs*
ESM	n x 325	"X"	Lin *et al.* (2022) *BioRxiv*

Not machine interpretable

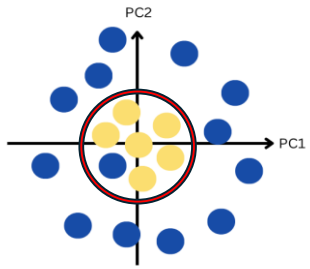
130,048 dimensions

2

```
>AAACCTGAGACCGGAT-6_VH
QVQLVQSGAEVKKPGASVKVSCKASGYTFTVFYIFWVRQA
PGQGPEWMGWINPNSGGTSYAQNFQGRVTMTRDTSVSTAY
MELSRLTSDDTAVYFCARGRRGLITEFDYWGQGTLVTVSS

>AAACCTGAGACCGGAT-6_VL
DIVMTQSPDSLAVSLGERATINCKSSQSVLDSSNNKNYLA
WYQQKPGQPPNLLIYWASTREYGVPDRFSGSGSGTDFTLT
ISSLQAEDVAVYYCQQYSSTPFTFGQGTKLEIK
```

```
[[-1.21275151  0.09233432 -0.28687406 ... -1.16596079  0.45893055
  1.03439796]
 [-1.37555611 -0.77200162 -1.6547246  ... -1.01659608 -0.25621107
  0.13018671]
 [-0.80599487  0.00688398 -0.39625698 ...  0.08043469 -0.37498006
 -0.36962718]
 ...
 [-1.55740356  0.66125804 -0.76089978 ... -0.27084541  0.02708428
 -0.73094225]
 [-1.66937947 -1.42654777 -1.21142054 ... -0.84660465  0.46885425
 -0.09832135]
 [-0.68128574  0.96581882 -0.21007441 ... -1.01042747 -0.05926855
  0.09461173]]
```



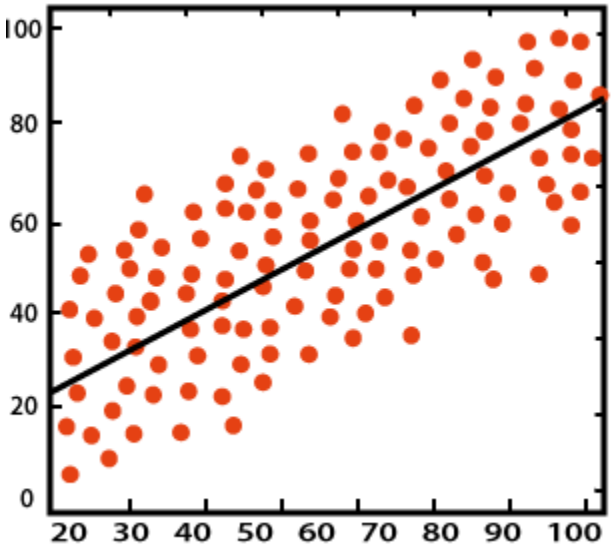
Antibody Sequences → Encoded Sequences → Encoded Sequences with dimensionality reduction

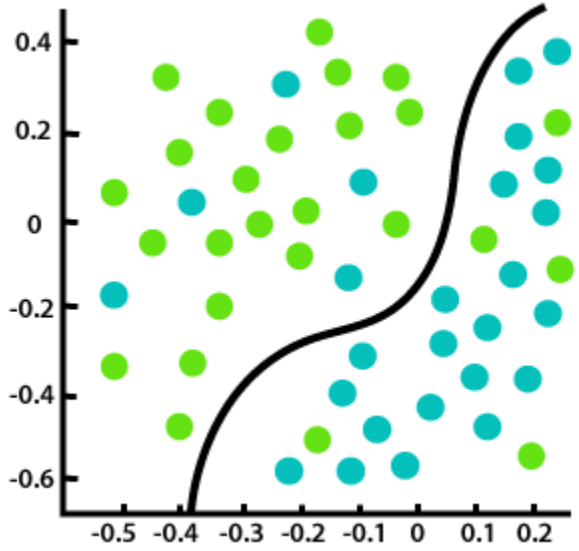
Bepler and Berger (2022) *Cell*

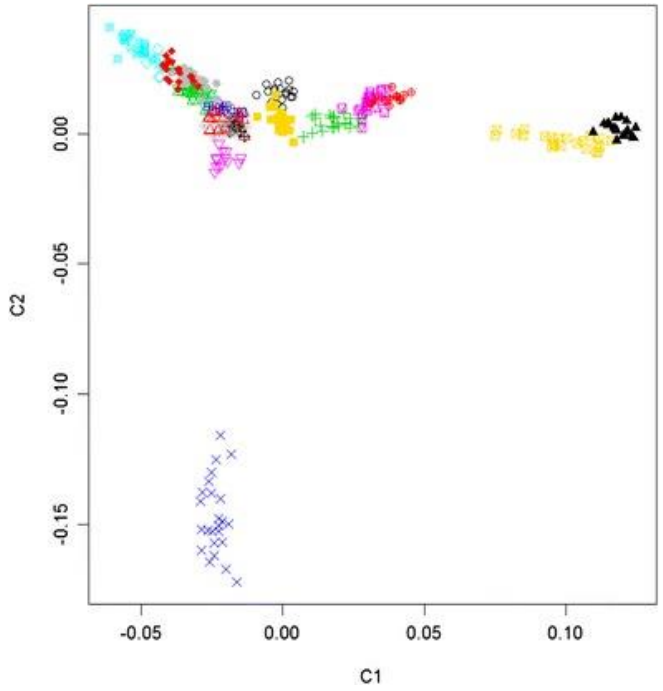
Machine Learning

Classification

Regression Supervised Unsupervised





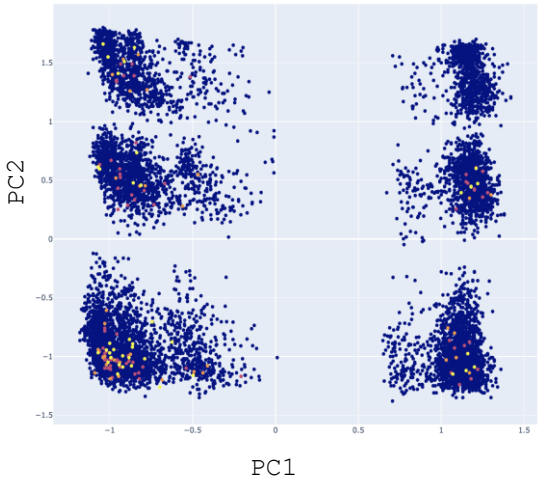


Unsupervised Learning

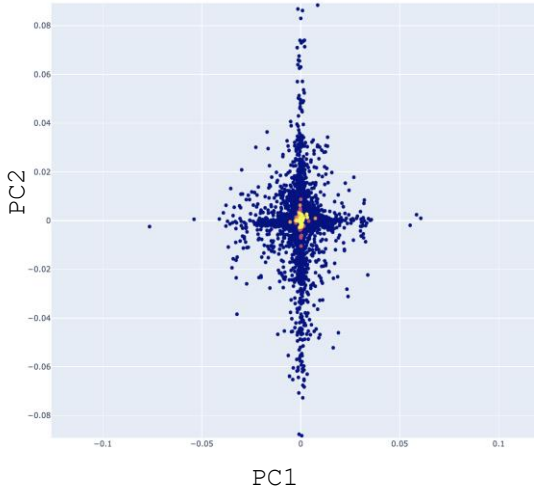
Training Data

- 10000 library human antibodies
- 31 approved human mAbs
- 77 discontinued human mAbs
- 35 human mAbs in clinical trials

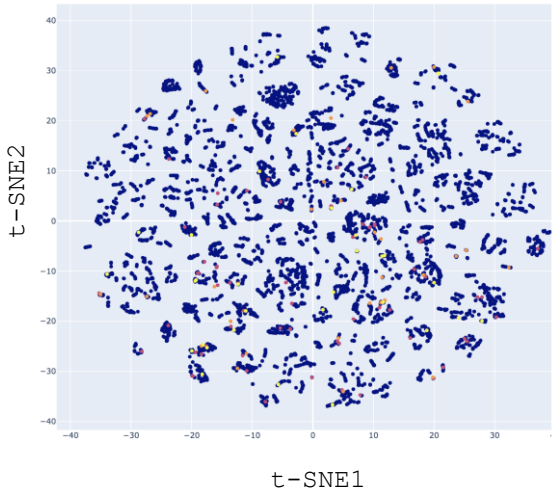
PCA



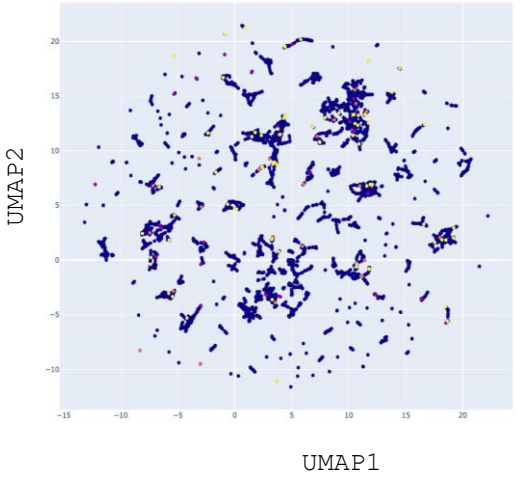
Kernel PCA

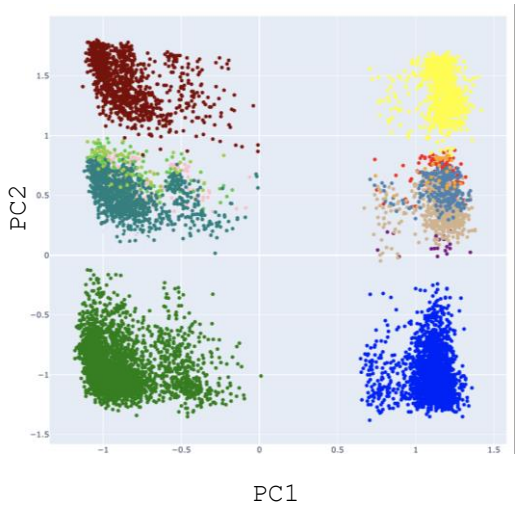


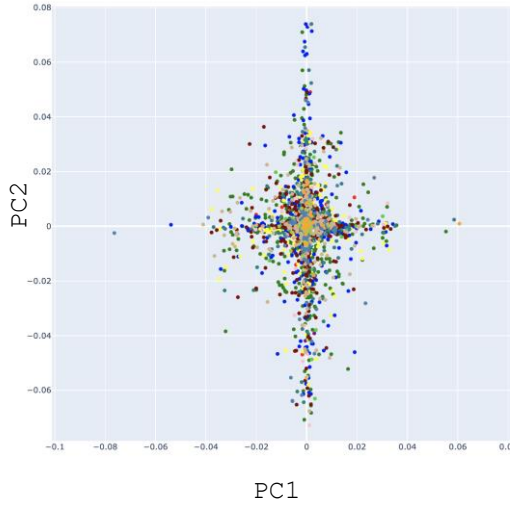
t-SNE

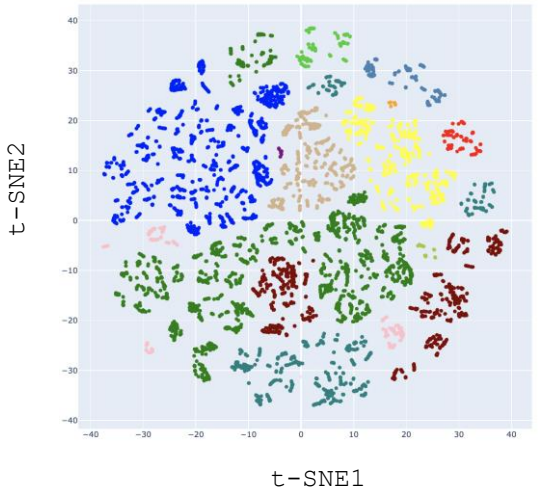


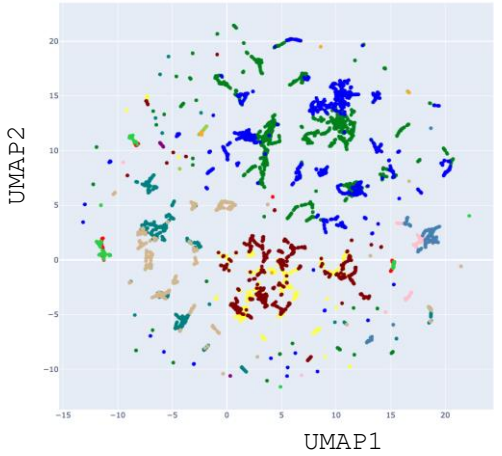
UMAP











Regression Tasks

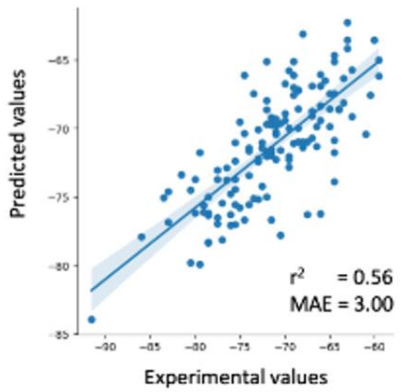
Training Data

- Experimental values of 136 clinical stage mAbs from Jain *et al.* (2017, 2023) *PNAS*

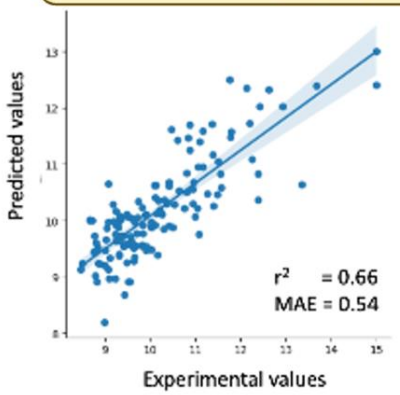
Developability Properties

- Stability
- Hydrophobicity
- Solubility
- Clearance
- Non-specific Binding

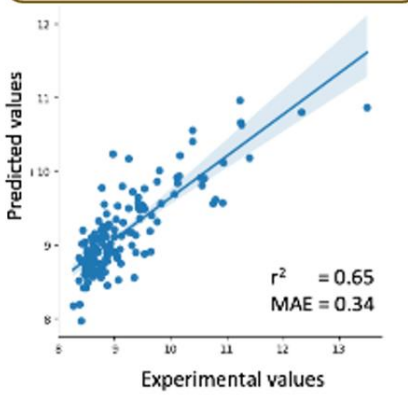
Melting Temperature



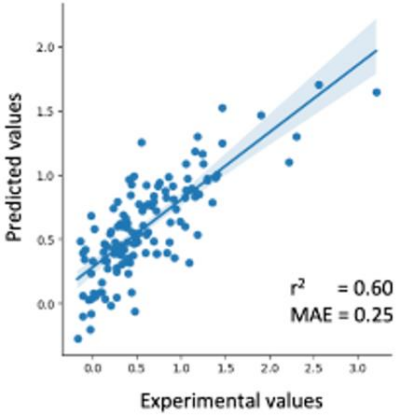
Hydrophobic Interaction Chromatography



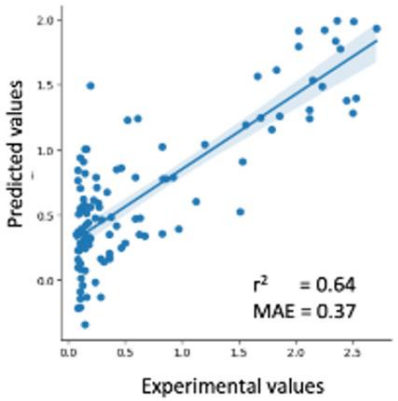
Cross-Interaction Chromatography



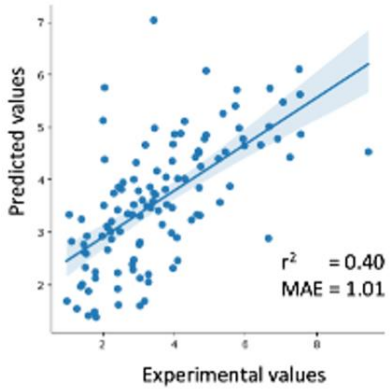
FcRn Retention Time



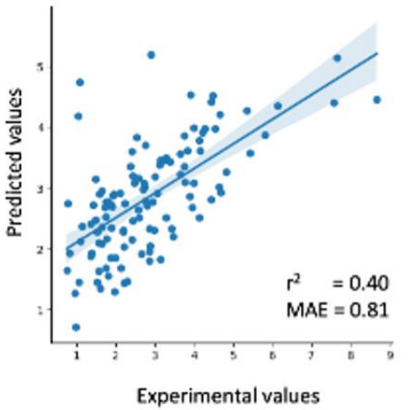
Binding to 2,4-Dinitrophenol



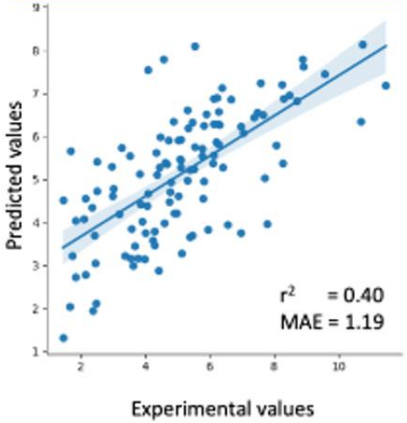
Binding to FVIII



Binding to C3



Binding to LysM

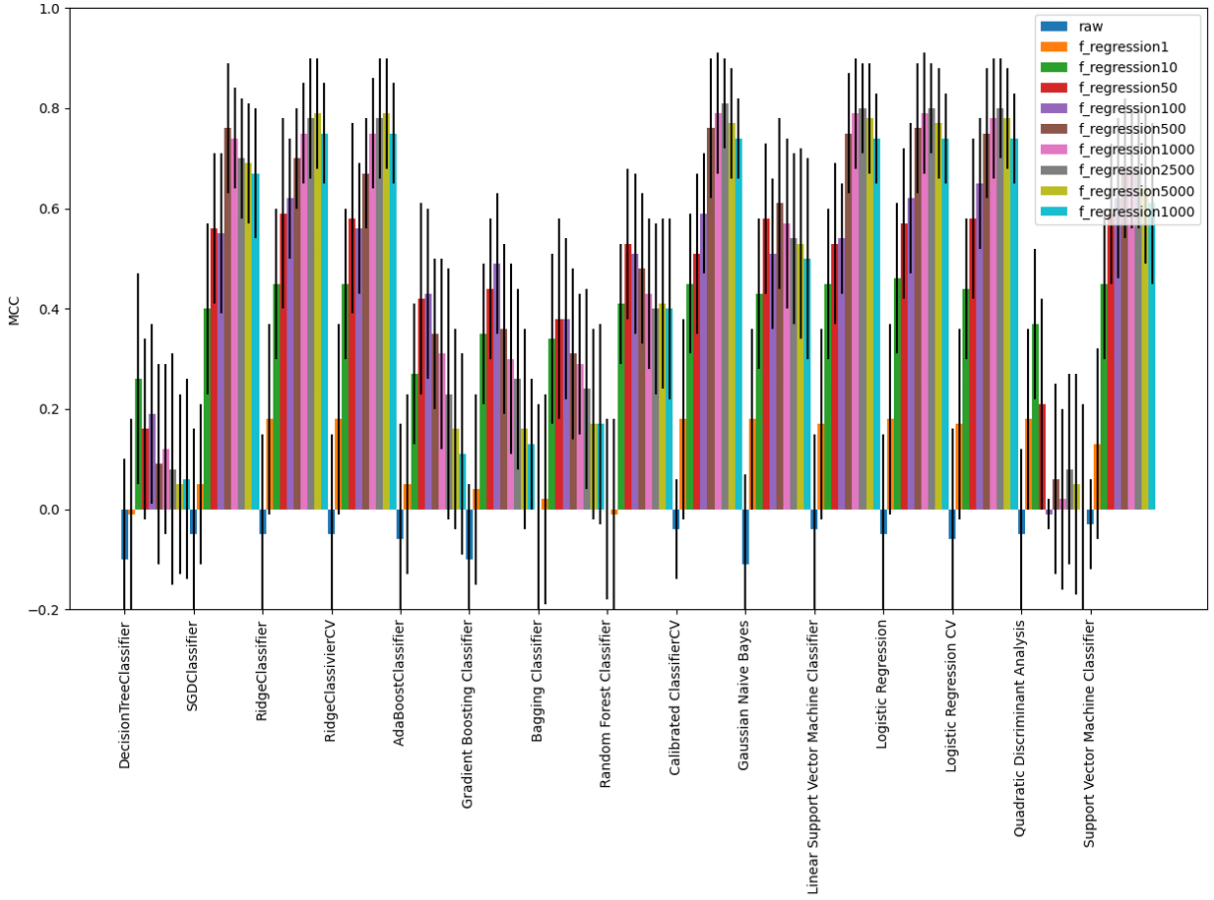


Supervised Learning

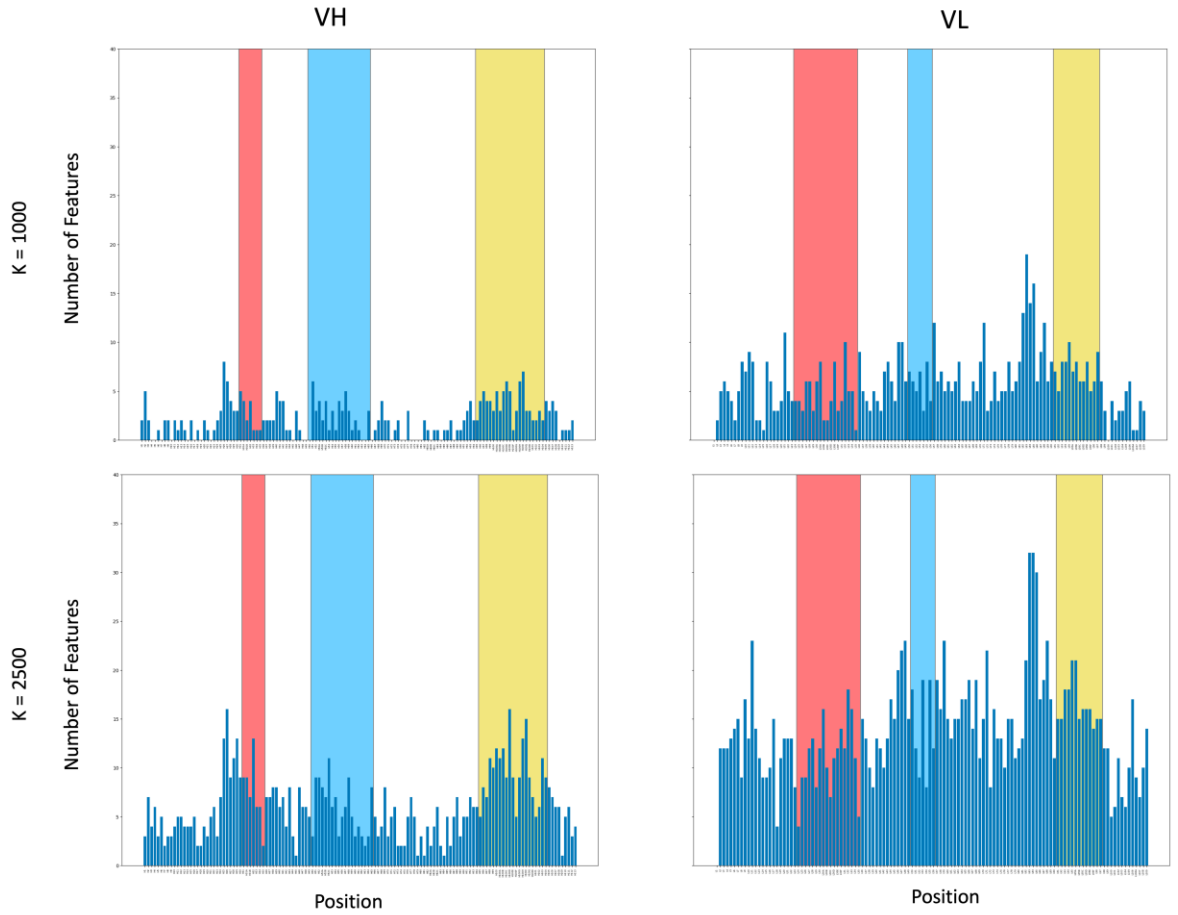
Training Data

- 115 Approved mAbs
- 145 Discontinued mAbs

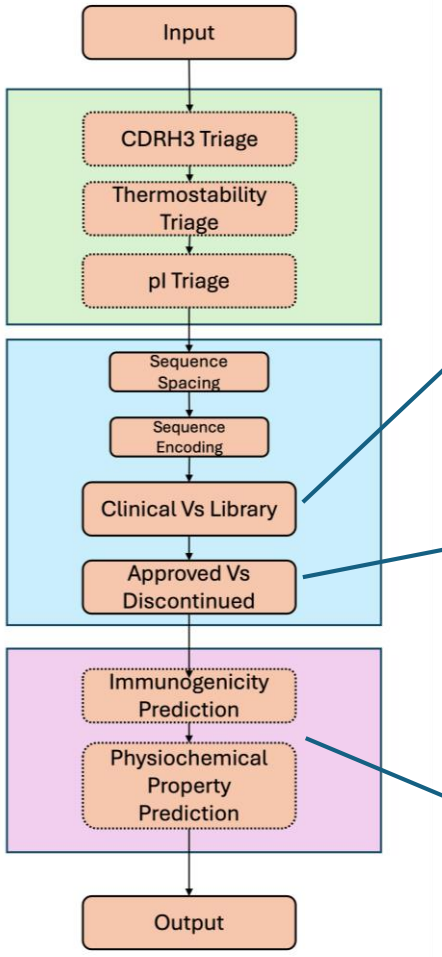
Binary Classifiers

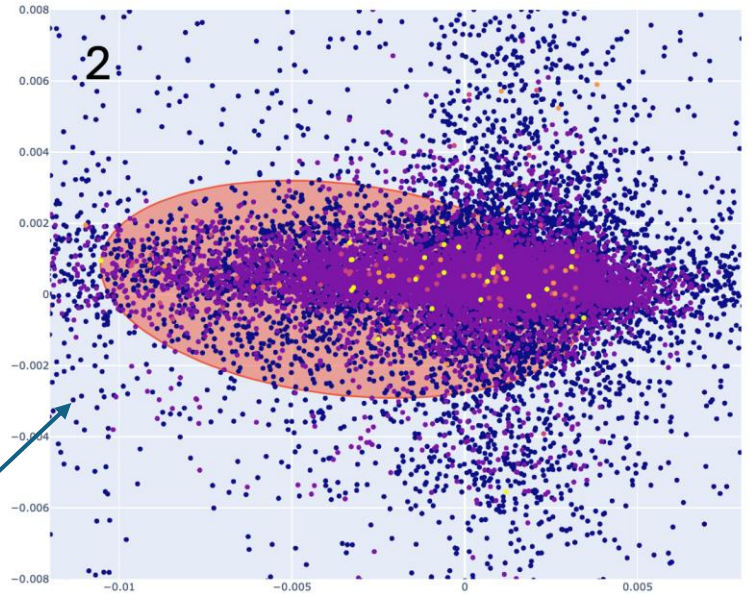


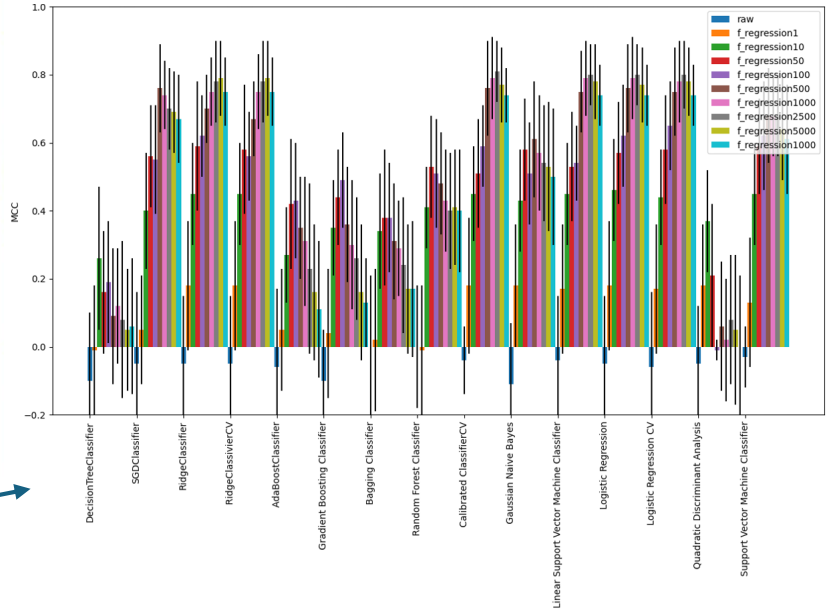
Location of Features

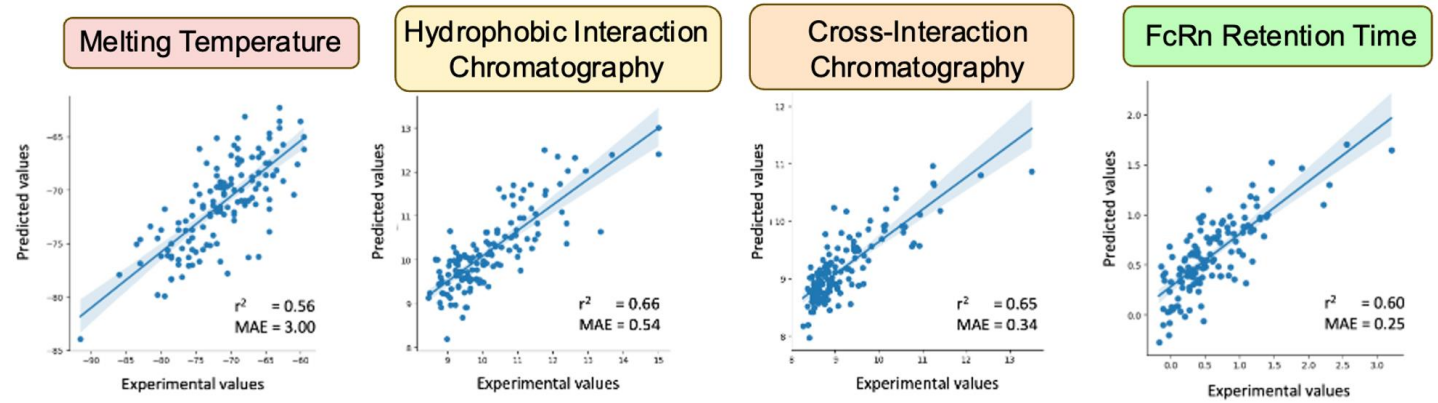


Outcomes


Input
CDRH3 Triage
Thermostability Triage
pI Triage
Sequence Spacing
Sequence Encoding
Clinical Vs Library
Approved Vs Discontinued
Immunogenicity Prediction
Physiochemical Property Prediction
Output


2


raw
f_regression1
f_regression10
f_regression50
f_regression100
f_regression500
f_regression1000
f_regression2500
f_regression5000
f_regression1000
MCC
DecisionTreeClassifier
SGDClassifier
RidgeClassifier
RidgeClassivierCV
AdaBoostClassifier
Gradient Boosting Classifier
Bagging Classifier
Random Forest Classifier
Calibrated ClassifierCV
Gaussian Naive Bayes
Linear Support Vector Machine Classifier
Logistic Regression
Logistic Regression CV
Quadratic Discriminant Analysis
Support Vector Machine Classifier


Melting Temperature
Hydrophobic Interaction Chromatography
Cross-Interaction Chromatography
FcRn Retention Time
Predicted values
Experimental values
r² = 0.56
MAE = 3.00
r² = 0.66
MAE = 0.54
r² = 0.65
MAE = 0.34
r² = 0.60
MAE = 0.25

After the Phl

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 July 2004 (08.07.2004)

PCT

(10) International Publication Number
WO 2004/056312 A2

(51) International Patent Classification[7]: A61K

(21) International Application Number:
PCT/US2003/040426

(22) International Filing Date:
16 December 2003 (16.12.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/434,115 16 December 2002 (16.12.2002) US
60/526,163 1 December 2003 (01.12.2003) US

(71) Applicant *(for all designated States except US)*: **GENENTECH, INC.** [US/US]; 1 DNA Way, SOUTH SAN FRANCISCO, CA 94080 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ADAMS, Camellia W.** [US/US]; 116C Flynn Avenue, Mountain View, CA 94043 (US). **CHAN, Andrew C.** [US/US]; 1201 Cloud Avenue, Menlo Park, CA 94025 (US). **CROWLEY, Craig W.** [US/US]; 151 Durazno Way, Portola Valley, CA 94028 (US). **LOWMAN, Henry B.** [US/US]; 400 San Juan Avenue, P. O. Box 2556, El Granada, CA 94018 (US). **NAKAMURA, Gerald R.** [US/US]; 1529 Portola Drive, San Francisco, CA 94127 (US). **PRESTA, Leonard G.** [US/US]; 1900 Gough Street, #206, San Francisco, CA 94109 (US).

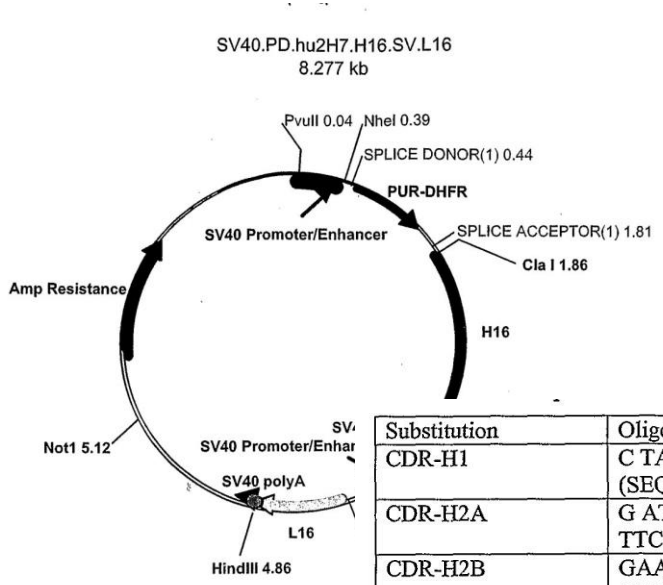
(74) Agent: **TAN, Lee K.**; c/o GENENTECH, INC., MS 49, 1 DNA Way, South San Francisco, CA 94080-4990 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (BW, GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.



Substitution	Oligonucleotide sequence
CDR-H1	C TAC ACC TTC ACG AGC TAT AAC ATG CAC TGG GTC CG (SEQ ID NO.)
CDR-H2A	G ATT AAT CCT GAC AAC GGC GAC ACG AGC TAT AAC CAG AAG TTC AAG GGC CG (SEQ ID NO. _)
CDR-H2B	GAA TGG GTT GCA GCG ATC TAT CCT GGC AAC GGC GAC AC (SEQ ID NO. _)
CDR-H3	AT TAT TGT GCT CGA GTG GTC TAC TAT AGC AAC AGC TAC TGG TAC TTC GAC GTC TGG GGT CAA GGA (SEQ ID NO. _)
CDR-L1	C TGC ACA GCC AGC TCT TCT GTC AGC TAT ATG CAT TG (SEQ ID NO. _)
CDR-L2	AA CTA CTG ATT TAC GCT CCA TCG AAC CTC GCG TCT GGA GTC C (SEQ ID NO. _)
CDR-L3	TAT TAC TGT CAA CAG TGG AGC TTC AAT CCG CCC ACA TTT GGA

WHAT IS CLAIMED IS:

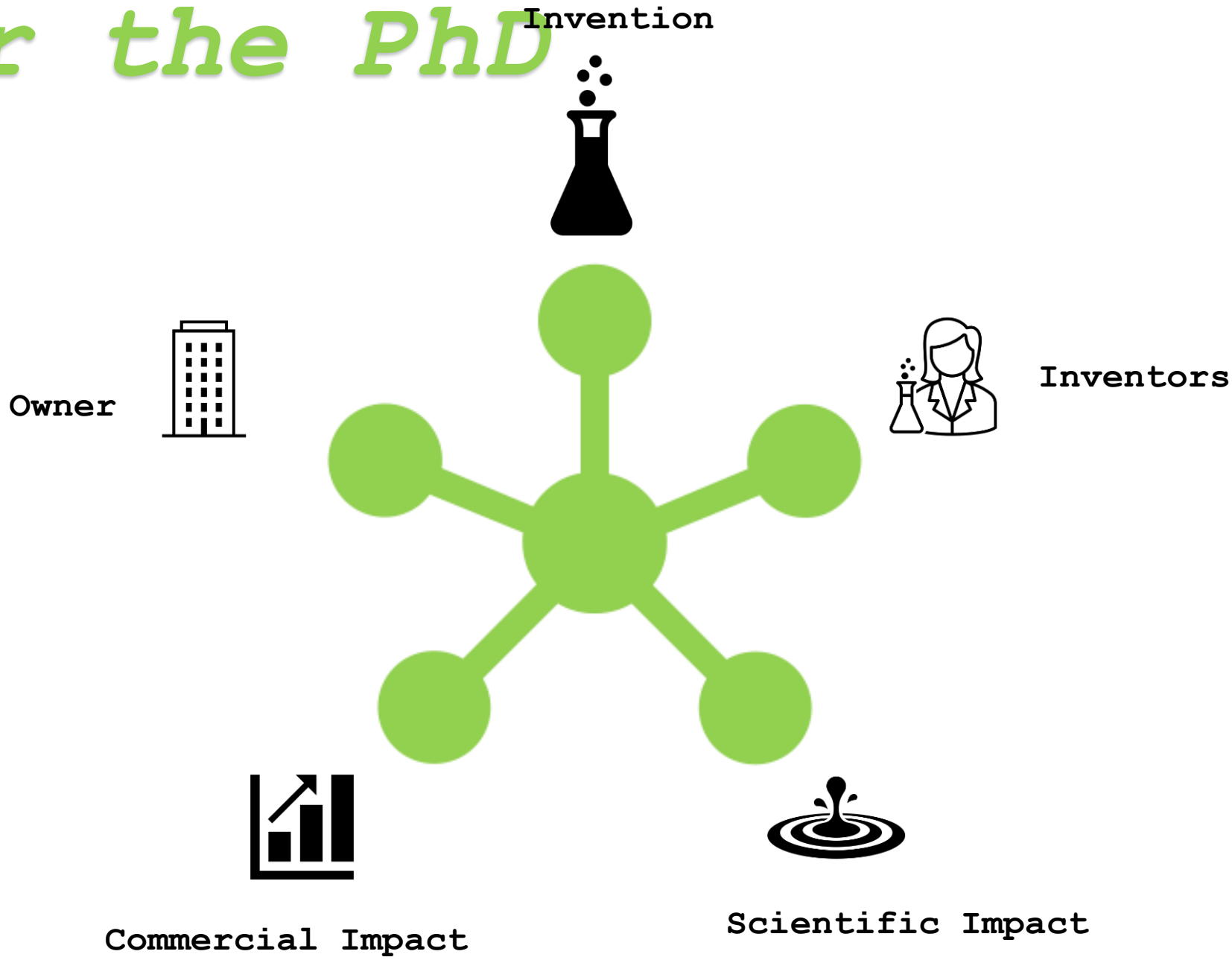
1. A humanized antibody that binds human CD20, or an antigen-binding fragment thereof, wherein the antibody is effective to deplete primate B cells in vivo, the antibody comprising in the H chain Variable region (V_H) at least a CDR3 sequence of SEQ ID NO. 12 from an anti-human CD20 antibody and substantially the human consensus framework (FR) residues of human heavy chain subgroup III (V_HIII).

2. The antibody of claim 1, further comprising the H chain CDR1 sequence of SEQ ID NO. 10 and CDR2 sequence of SEQ ID NO. 11.

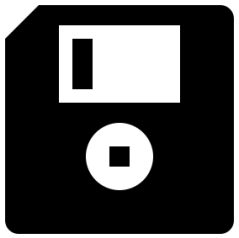
3. The antibody of claim 2, further comprising the L chain CDR1 sequence of SEQ ID NO. 4, CDR2 sequence of SEQ ID NO. 5, CDR3 sequence of SEQ ID NO. 6 and substantially the human consensus framework (FR) residues of human light chain κ subgroup I (VκI).

4. The antibody of the preceding claims, comprising the V_H sequence of SEQ ID NO.8 (v16, as shown in FIG. 1B).

After the PhD



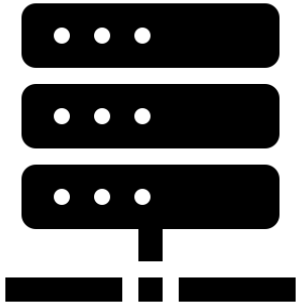
How to get the most out of AI



High quality data

Suitable Problem

Programming skills



Compute

Explainability

Non obvious insights

Thanks for listening!