Supporting Information

Intrinsic Physicochemical Profile of Marketed Antibody-based Biotherapeutics

Lucky Ahmed[1], Priyanka Gupta, Kyle Martin, Justin M. Scheer[2], Andrew E. Nixon and Sandeep Kumar*

Biotherapeutics Discovery, Boehringer Ingelheim, Ridgefield, USA
[1]Current address: Just Evotec Biologics, 401 Terry Avenue North, Seattle, WA 98109
[2]Current address: Gene Therapy and Gene Delivery Platforms, The Janssen Pharmaceutical Companies of Johnson & Johnson, 260 E Grand Ave, South San Francisco, CA 94080

* Corresponding author: Sandeep Kumar

Email: sandeep_2.kumar@boehringer-ingelheim.com

Contents

Materials and Methods

Sequence collection
Amino acid sequences of 77 antibody-based biotherapeutics currently available in the market were collected and verified using five different databases, namely, IMGT (http://www.imgt.org/mAb-DB/), Drug Bank (https://www.drugbank.ca/), NIH (https://drugs.ncats.io/substances), SciFinder (https://scifinder-n.cas.org) and INN (https://www.who.int/medicines/services/inn/en/). We have also collected 271 antibody sequences currently in clinical development (phase I – III). The initial set of 242 clinic stage (CST) antibody sequences was collected from Raybould et al (1). We found and excluded 70 marketed antibody sequences in the 242 CST dataset. The remaining 172 CST antibody sequences were combined with 99 additional CST sequences obtained by searching the IMGT database (http://www.imgt.org/mAb-DB/). Names of the 271 CST antibodies used in this study are provided in Dataset S1. Furthermore, 14037 human sequences from next-generation sequencing (NGS) repertoires were also collected from Raybould et al (1). In addition to these datasets, 3120 of our internal hit sequences from several different antibody generation campaigns, involving humanized mice and phage display of human germline sequences, were also used in this work.

Germline Analysis
The germline analysis was performed for 79 Fv sequences using the IgBLAST (from NCBI, updated on 2018, https://ftp.ncbi.nih.gov/blast/executables/igblast/release/).

Homology Modeling
Molecular models of 79 Fv regions found in 77 marketed biotherapeutics were built using automated homology modeling in Molecular Operating Environment (MOE) (2) by Chemical Computing Group (www.chemcomp.com). Antibody modeler with default Amber10: EHT force field, internal and external dielectric values of 4 and 80, the non-bonded cutoff distances of 10 and 12 Å and the Born solvation was used to build these homology-based models. After building the homology models, the C-termini of the light and heavy chains were amidated to neutralize charges on them. The capped models were prepared for energy minimizations by utilizing the structure preparation to remove any errors and protonate the models at pH 7.0 and 0.1 M salt. The prepared structures were energy minimized to root man square gradient (RMSG) below 0.00001 kcal/mol/Å^2. Fv region models for 271 clinical stage antibodies, 14037 human antibodies, and 3120 internal hits were built using the same protocol.

Protein Properties Calculation
For each Fv region, we calculated protein properties using both the energy minimized models and conformer ensembles generated using LowModeMD as implemented in MOE 2018 (2). In the LowModeMD calculations, up to 50 conformers were generated using the energy minimized model of each of the 79 Fvs. The framework backbone atoms were restrained to 0.25 Å of their modeled positions and the C$^\alpha$ atoms of the HCDR3 loops were restrained to 2Å. The C$^\alpha$ atoms of the other CDRs were restrained to 1Å. The protein property calculations were repeated for each conformer and the average values of these properties were taken for further analysis. In addition to descriptors computed in MOE protein properties calculations, two new descriptors were also devised, namely, the ratio of surface areas of positively and negatively charged patches to that of hydrophobic patches (RP), and the ratio of dipole

moment to hydrophobic moment (RM). The ratio of surfaces areas of charged patches to hydrophobic patches is calculated as:

$$RP = \frac{Surface\ area\ of\ Positively\ charged\ patches + Surface\ area\ of\ Negatively\ charged\ patches}{Surface\ area\ of\ Hydrphobic\ patches} \quad \ldots\ldots (1)$$

The ratio of dipole moment to the hydrophobic moment is been calculated as:

$$RM = \frac{dipole\ moment(\boldsymbol{\mu_D})}{hydrphobic\ moment(\boldsymbol{\mu_H})} \quad \ldots\ldots (2)$$

Data Analysis

Each descriptor was examined for variance among its values over the dataset of 79 Fv regions. A total of 24 different descriptors, that show significant variations among different Fv models, were obtained from protein properties calculations and analyzed in several stages as follows. At first, Pearson correlation coefficients *r* were computed among all descriptors via Graph Pad Prism 8 (3). Then the computed *r* - values were utilized to make the clusters as described below. Because different protein properties can represent similar structural or physicochemical features of proteins, hierarchical clustering was used to decipher relationships among different descriptors. In this case, python pandas packages (4, 5) were applied to plot a hierarchical dendrogram to identify clusters and non-redundant descriptors. In this case, a Rule of Thumb (6) was applied to find non-redundant descriptors:

A statistically significant linear correlation may exist between descriptors x and y, if

$$|r_{xy}| \geq \frac{2}{\sqrt{n}} \quad \ldots\ldots (3)$$

Where r is the correlation coefficient and n is the sample size.

In this study, the sample size is n = 79, therefore $r_{xy} = 2/\sqrt{79} \sim 0.23$. Hence, a cutoff value of 0.23 was used to identify the non-correlated clusters of descriptors. All descriptors with ($r \geq |0.23|$) were clustered together in a hierarchical dendrogram shown in Figure 1B. In the next step, the descriptors in each cluster were examined for similarity of their physicochemical meaning and if a cluster happens to contain descriptors with significantly divergent physicochemical meanings, then one descriptor from each divergent subcluster was selected. For example, the second cluster in Figure 1B contains RM (ratio of dipole moment to hydrophobic moment), Dipole Moment, Hydrophobic Moment, Average Eint_V_L: V_H and BSA_V_L: V_H. Two descriptors, namely, RM and BSA_V_L: V_H were selected (r = 0.26). This exercise has yielded five non-redundant descriptors, namely, BSA_V_L: VH, pI_{Fv}_3D, RP, RM and Avg_hydrophobic imbalance (Avg_HI). This procedure was repeated four times by setting aside 10 randomly selected Fvs and similar results were obtained (Figures S3A-D).

Statistical summary (mean, standard deviation and range) of these five non-redundant descriptors profile structure-based physicochemical attributes of the Fv regions of 77 marketed antibody-based biotherapeutics are shown in Table 2 along with several subsets for which the calculations were repeated in the same way as the full dataset. This profile was used to compare the physicochemical properties of the Fv regions of 271 CST, 14037 human and 3120 internal hits with those of marketed biotherapeutics by computing Z-scores for all descriptors:

$$Z_i = (x_{ij} - \mu_i)/\sigma_i \quad \ldots\ldots (4)$$

Where i = 1 to 5 for five non-redundant descriptors, namely, BSA_V_L: VH, pI_{Fv}_3D, RP, RM and HI. x_{ij} is the value of i^{th} descriptor for the Fv region of the j^{th} sequence in collections of 271 CST, 14037 human or 3120 internal hit antibodies. μ_i and σ_i are mean and standard deviation values of the ith descriptor for 77 marketed antibody-based biotherapeutics. Z-scores of the individual descriptors can identify the structural features of an Fv region that need to be optimized via protein engineering to lower its Z-distance. The individual Z-scores were also used to flag the Fv regions in different antibody datasets if their magnitudes exceeded 1.96.

The calculated Z-scores for a given Fv region were combined into a single metric called Z-distance (Z-distance) using the following equation:

$$Z - distance = \sqrt[2]{\sum_{i=1,5} Z_i^2} \quad \text{..... (5)}$$

Z-distance measures the similarity of the physicochemical properties of an Fv region with those of the marketed biotherapeutics.

References

1. M. I. Raybould *et al.*, Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences* **116**, 4025-4030 (2019).
2. C. C. G. Inc. (2018) Molecular operating environment (MOE). (Chemical Computing Group Inc 1010 Sherbooke St. West, Suite# 910, Montreal).
3. M. L. Swift, GraphPad prism, data analysis, and scientific graphing. *Journal of chemical information and computer sciences* **37**, 411-412 (1997).
4. W. McKinney, "Python for data analysis: Data wrangling with Pandas, NumPy, and IPython" (O'Reilly Media, 2012).
5. P. Virtanen *et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods* **17**, 261-272 (2020).
6. T. C. Krehbiel, Correlation Coefficient Rule of Thumb. *Decision Sciences Journal of Innovative Education* **2**, 97-100 (2004).

Supporting Figures

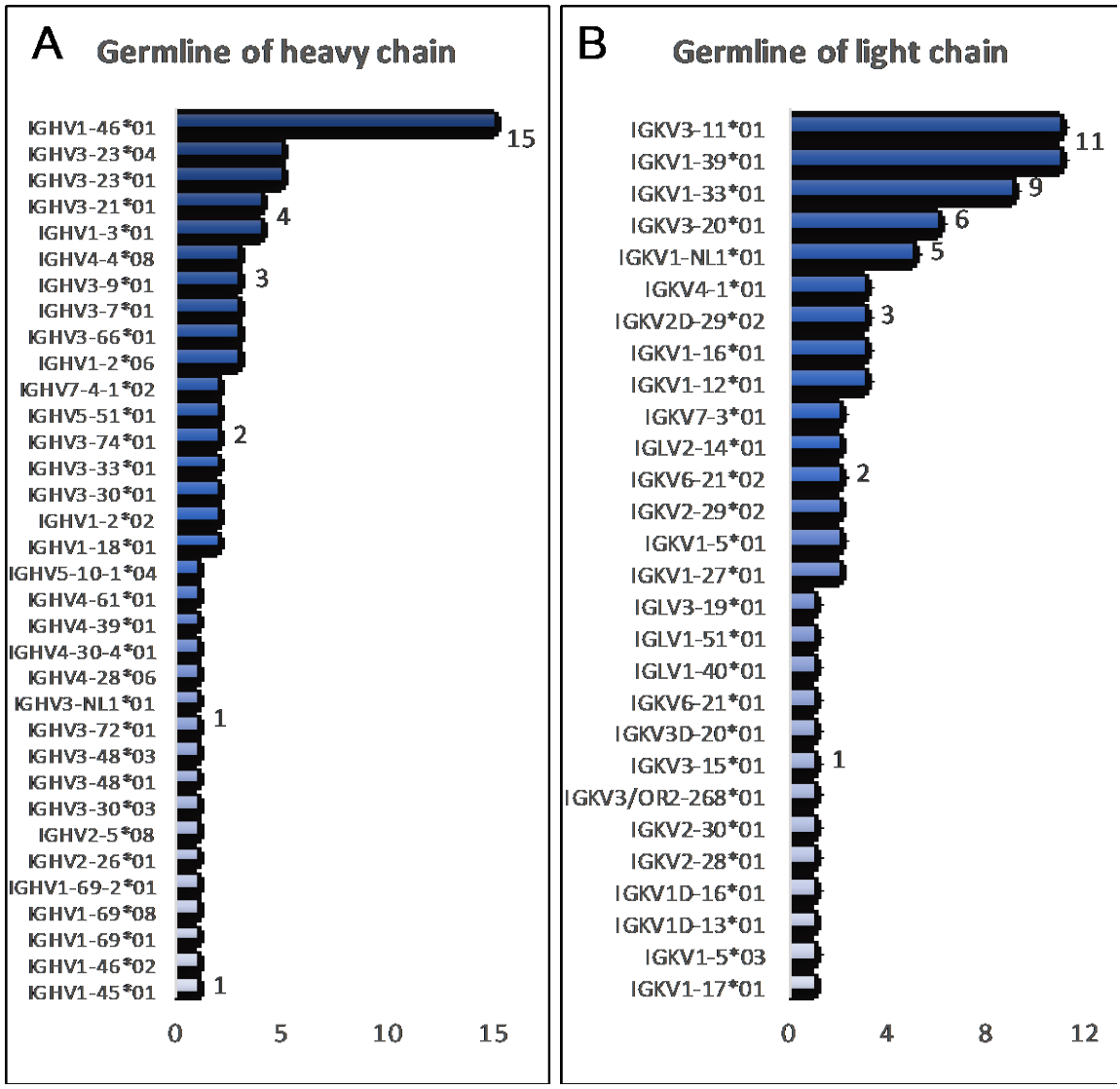

Figure S1: (A) Germline analysis of 79 heavy chains from 77 marketed antibodies. (B) Germline analysis of 79 light chains from 77 marketed antibodies.

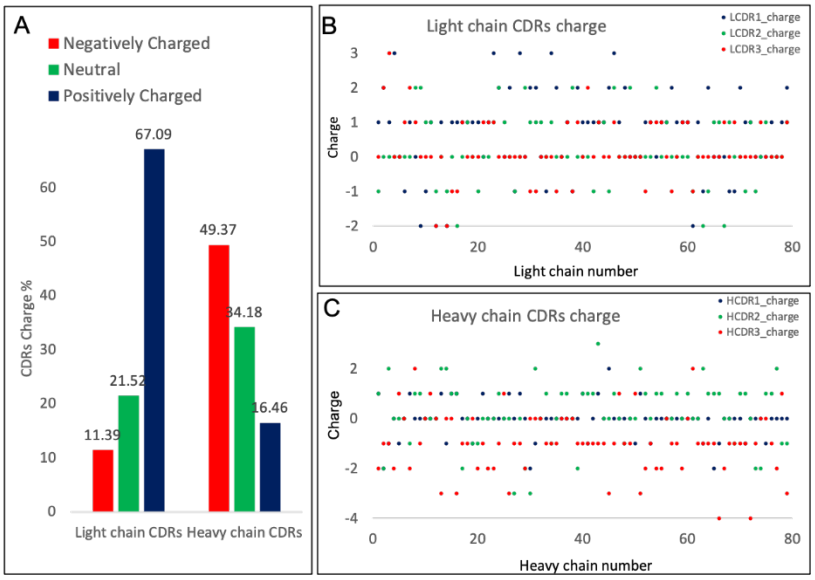

Figure S2: (A) Histograms showing CDR charge diversity in 77 marketed antibody-based biotherapeutics demonstrate an asymmetry in the CDR charge among them. (**B**) Charge diversity in light chain CDRs with most of them being positively charged (67.1%). (**C**) Charge diversity in heavy chain CDRs with approximately half of them being negatively charged (49.4%).

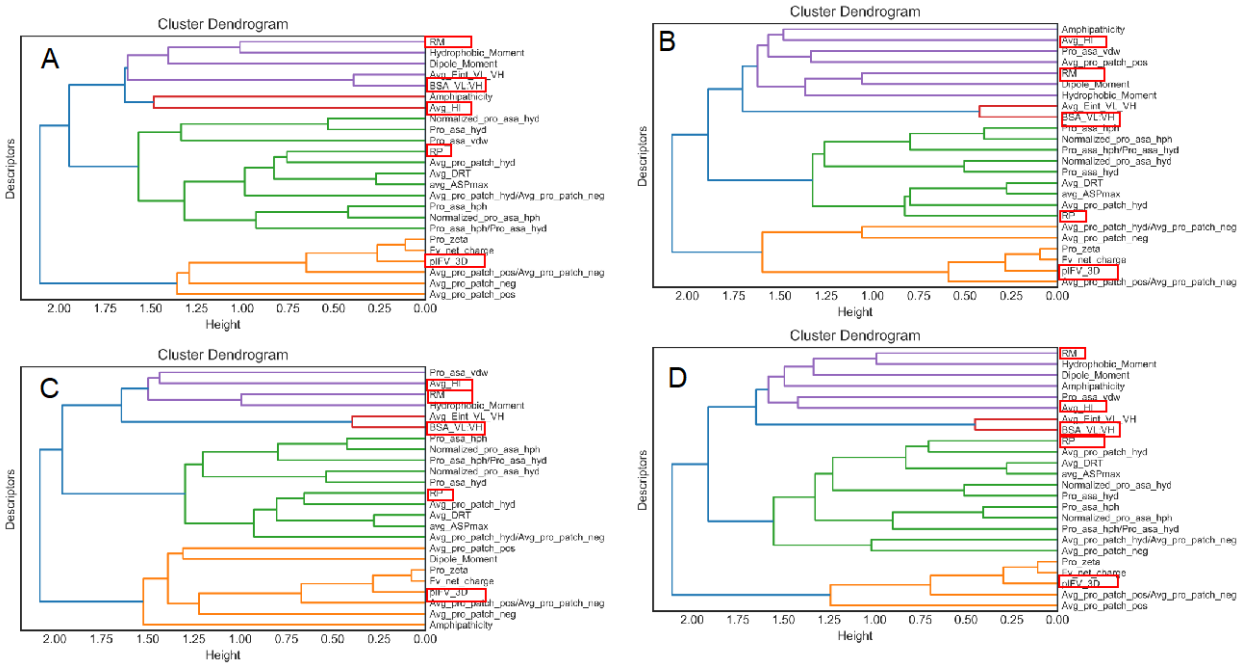

Figure S3: Clustering of partial list of 69 Fvs, 10 Fvs are removed randomly from the main list of 79 Fvs. Four (A, B, C & D) different partial datasets were prepared for clustering. Five non-redundant descriptors are highlight in red boxes.

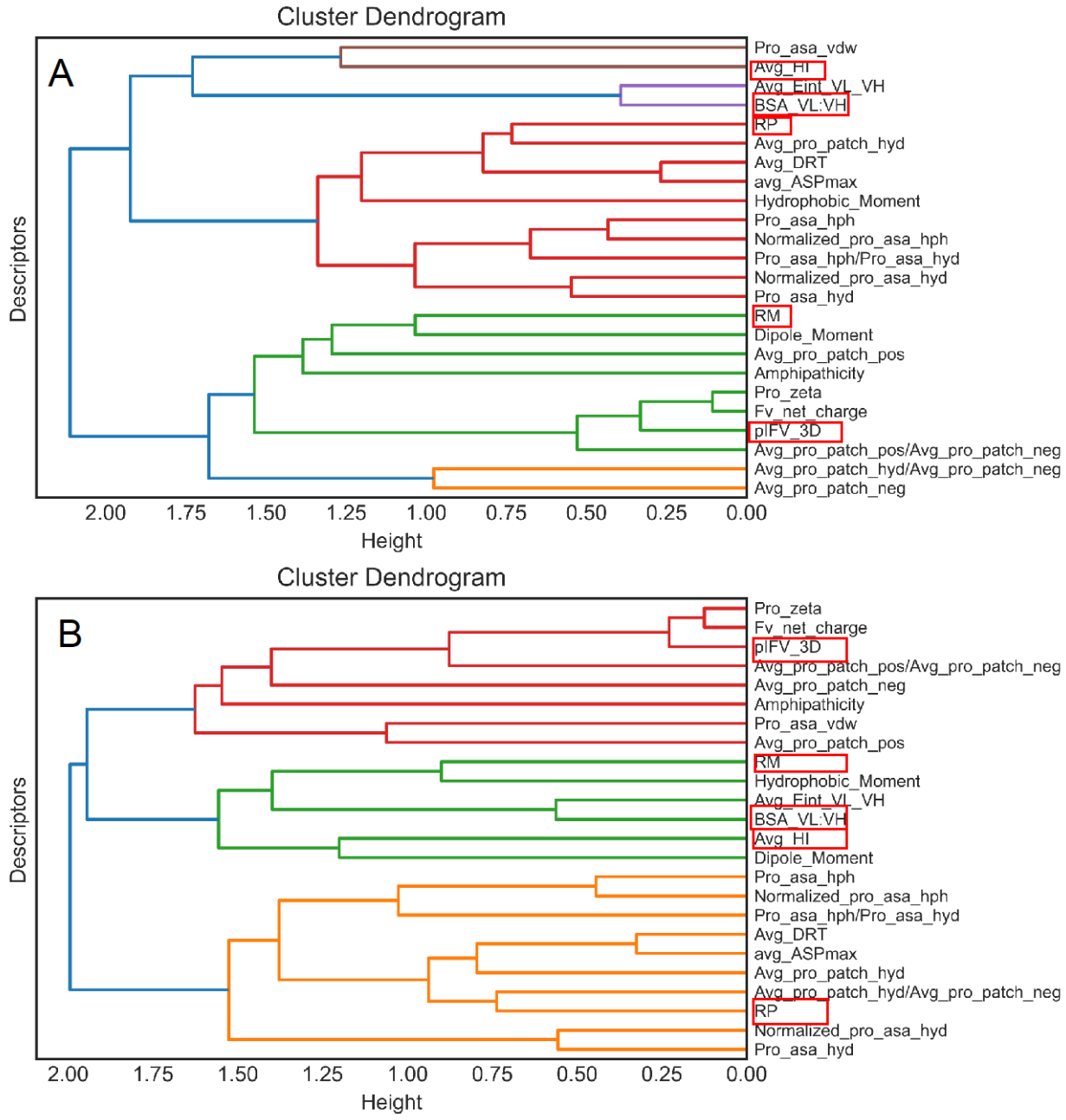

Figure S4: (A) Clustering of calculated descriptors from homology-based models of 36 Fvs from biologic medicines approved in 2015 and after. (B) Clustering of calculated descriptors from homology-based models of 43 Fvs from biologic medicines approved till 2014. Five non-redundant descriptors are highlight in red boxes.

Supporting Datasets & Tables

Dataset S1: See the Excel Sheet named **DatasetS1.xlsx** for sequences and datasets for all approved antibodies, 271 CST antibodies and 14037 human antibodies.

Dataset S2: See the Excel Sheet named **DatasetS2.xlsx** for data on different partial sets analyzed in this work.

Table S1: Average and standard deviation of different CDRs lengths and hydrophobic residues in the CDRs for 77 marketed antibody-based biologic medicines.

CDRs	Length (Average ± std)	Hydrophobic residues (#) (Average ± std)	Hydrophobic residues (%) (Average ± std)
HDR1	10±1	6±1	58±8
HDR2	17±1	9±1	51±8
HCDR3	11±3	7±2	66±13
LCDR1	12±2	5 ± 1	43±7
LCDR2	7±0	3± 1	38±12
LCDR3	9±1	4±1	42±12

Table S2: Calculated five descriptors of G6 and CNTO antibodies and their variants.

Antibody	BSA_V_L:V_H (Å^2)	pI_{Fv}_3D	RM(μ_D/μ_H) (D)	RP	Avg_HI
G6	885	6.4	0.5	1.1	1.3
G6_LC_S52R	897	7.5	0.3	1.1	1.4
CNTO	757	4.3	1.2	0.9	0.3
CNTO_HC_F99A	764	4.5	1.2	1.3	0.5
CNTO_HC_H100A	749	4.2	1.3	0.9	0.2
CNTO_HC_W100aA	751	4.3	1.2	1.1	0.6
CNTO_HC_F99A-H100A-W100aA	741	4.1	1.3	1.4	0.7