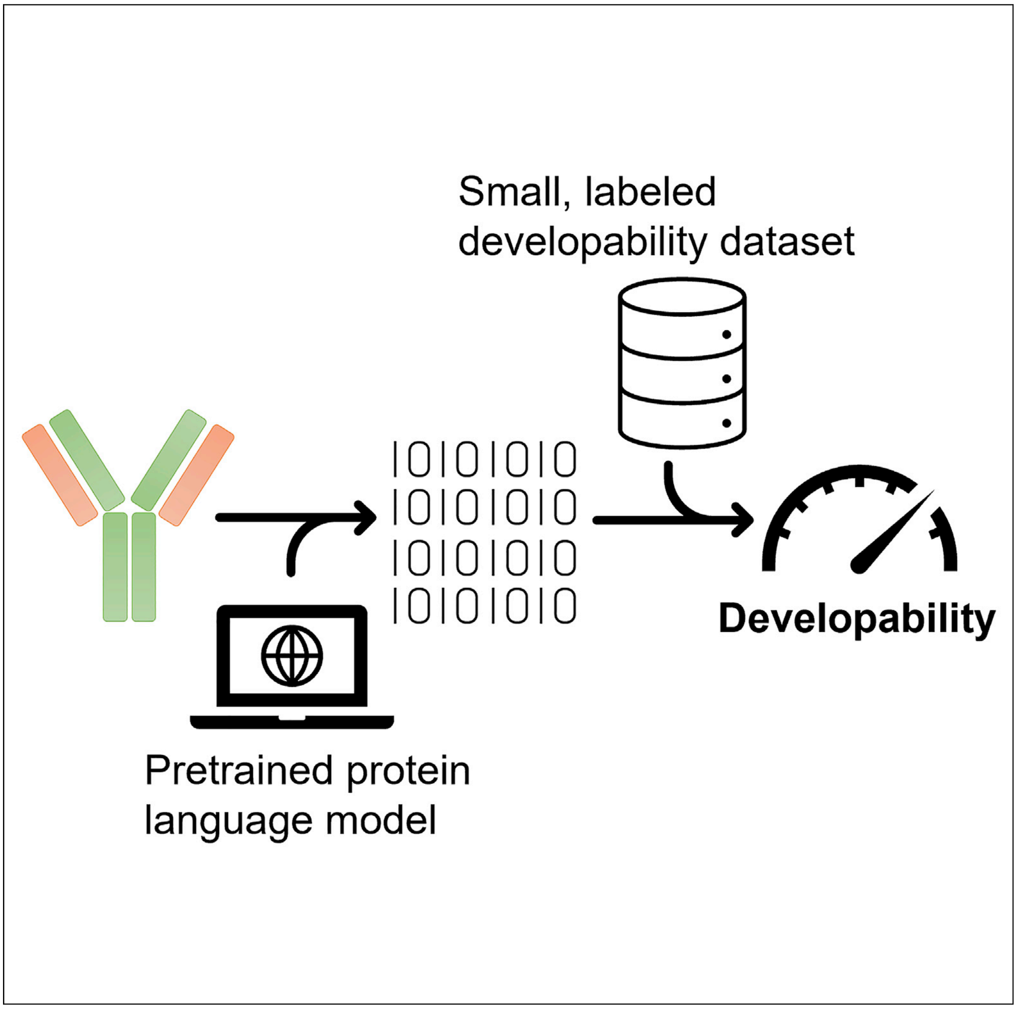
iScience

CellPress
OPEN ACCESS

Article

Antibody apparent solubility prediction from sequence by transfer learning



Jiangyan Feng, Min Jiang, James Shih, Qing Chai

feng_jiangyan@lilly.com (J.F.)
chai_qing_qc@lilly.com (Q.C.)

Highlights

Rapid and high-throughput antibody solubility prediction using sequence alone

Pretrained protein embeddings are biologically meaningful for antibodies

Transfer learning alleviates data scarcity for antibody developability prediction

Feng et al., iScience *25*, 105173
October 21, 2022 © 2022 The Author(s).
https://doi.org/10.1016/j.isci.2022.105173

iScience

CellPress OPEN ACCESS

Article

Antibody apparent solubility prediction from sequence by transfer learning

Jiangyan Feng,[1,*] Min Jiang,[2] James Shih,[1] and Qing Chai[1,3,*]

SUMMARY

Developing therapeutic monoclonal antibodies (mAbs) for the subcutaneous administration requires identifying mAbs with superior solubility that are amenable for high-concentration formulation. However, experimental screening is often material and labor intensive. Here, we present a strategy (named solPredict) that employs the embeddings from pretrained protein language modeling to predict the apparent solubility of mAbs in histidine (pH 6.0) buffer. A dataset of 220 diverse, in-house mAbs were used for model training and hyperparameter tuning through 5-fold cross validation. solPredict achieves high correlation with experimental solubility on an independent test set of 40 mAbs. Importantly, solPredict performs well for both IgG1 and IgG4 subclasses despite the distinct solubility behaviors. This approach eliminates the need of 3D structure modeling of mAbs, descriptor computation, and expert-crafted input features. The minimal computational expense of solPredict enables rapid, large-scale, and high-throughput screening of mAbs using sequence information alone during early antibody discovery.

INTRODUCTION

Therapeutic monoclonal antibodies (mAbs) represent the fastest growing class of therapeutics on the market, with around 100 antibody drugs approved to treat a wide spectrum of human diseases (Leavy, 2010), including cancer (Dean et al., 2021; Weiner et al., 2010), inflammatory, and autoimmune diseases (Chan and Carter, 2010). Subcutaneous injection has emerged to be the preferred delivery route of mAbs drug products especially in the treatment of chronic diseases, because they can be self-administered at home and therefore enhances patient adherence and compliance (Anselmo et al., 2019). Given limited injection volume (<2 mL) and high dose requirement (~500 mg), mAbs must be soluble enough to achieve high-concentration formulations (>100 mg/mL) (Kingsbury et al., 2020). Furthermore, mAbs must remain soluble at high concentrations during the manufacturing process which can cause protein precipitation. Therefore, superior solubility is vital for developing liquid formulation of therapeutic mAbs (Makowski et al., 2021; Shire et al., 2004; Wolf Pérez et al., 2022).

A practical hurdle is that poor solubility behavior often manifests at higher mAb concentrations (>50 mg/mL) (Chai et al., 2019). Early experimental screening is often challenged by the large number of antibody candidates and the limited preparation quality available (i.e. minute amounts, low concentrations, and low purity) (Chai et al., 2019; Wolf Pérez et al., 2019). *In silico* solubility prediction appears to be a convenient alternative owing to its capability of rapid high-throughput screening without material requirement (Han et al., 2022; Hebditch et al., 2017; Sormanni et al., 2015, 2017). Current computational approaches rely on molecular descriptors extracted either from protein sequence (sequence-based predictors (Hebditch et al., 2017; Sormanni et al., 2017)) or from structures (structure-based predictors (Chan et al., 2013; Han et al., 2022; Sormanni et al., 2015)). Sequence-based predictors often neglect tertiary structure information, which distinguishes poorly soluble residues driving protein folding from the ones that are exposed to the solvent and may elicit aggregation (Wolf Pérez et al., 2022). Structure-based tools can be used only when the structure or a high-quality model is available. This limits the throughput and application to large number of early-stage mAb candidates. Furthermore, some of the computational methods only output a binary classification (e.g. soluble/insoluble) (Hebditch et al., 2017; Smialowski et al., 2012; Trainor et al., 2017) instead of a numerical value.

The lack of quantitative solubility dataset of large, diverse mAbs at pharmaceutically relevant formulation further hinders the generalizability of computational predictors. Previous developability related work has

[1] BioTechnology Discovery Research, Eli Lilly Biotechnology Center, San Diego, CA 92121, USA

[2] Advanced Analytics and Data Sciences, Eli Lilly Corporate Center, Indianapolis, IN 46225, USA

[3] Lead contact

*Correspondence: feng_jiangyan@lilly.com (J.F.), chai_qing_qc@lilly.com (Q.C.)

https://doi.org/10.1016/j.isci.2022.105173

Table 1. Summary of solubility behavior of two control mAbs at H6

			H6 (PEG %)	
mAb	Subclass	Solubility behavior	mean	SD
mAb239 (control 1)	IgG1	Good solubility	35.81	0.72
mAb240 (control 2)	IgG4	Poor solubility	10.63	1.46

been performed with non-mAbs proteins (Hebditch et al., 2017), limited mAb datasets (Sormanni et al., 2015, 2017), closely related mAbs with varying mutations (Sormanni et al., 2015, 2017), or mAbs belonging to the same subclass (Sharma et al., 2014). Furthermore, mAb solubility is highly dependent on formulation condition (Chai et al., 2019). Histidine and pH 6.0 (H6) buffer system has emerged as a common buffer/pH system for mAb-based products, because at pH 6.0 chemical degradation of proteins is minimized which makes liquid formulations feasible (Kingsbury et al., 2020). To the best of our knowledge, there are no computational tools that can quantitatively predict the solubility at H6 condition for different subclasses of mAbs using sequence information alone.

A powerful approach is transfer learning which distills the knowledge learned from protein language models trained on very large unlabeled protein sequences and builds a downstream model supervised with limited data (Bepler and Berger, 2021). Previous studies have shown that using the representations from pretrained protein language models as input features to fine-tune a supervised model can improve a wide variety of protein-relevant tasks, including secondary structure prediction, contact prediction, and remote homology detection (Rao et al., 2019; Rives et al., 2021). However, such methods have not been fully explored to predict attributes that are essential to antibody developability.

Here, we propose an end-to-end sequence-to-function model, solPredict, to predict the quantitative solubility of mAbs at H6 using only antibody sequences. To develop a general predictor of antibody solubility, we constructed a large and diverse set of 260 in-house mAbs, consisting of 112 IgG1, 140 IgG4, and 8 IgG2. The quantitative solubility was measured at histidine pH 6.0 buffer condition using PEG-induced precipitation method due to the advantages of high-throughput screening and minimal material requirement (Chai et al., 2019). To overcome the limitation of expert-crafted descriptors, and the necessity to obtain high-quality 3D structures, we represented antibody sequences as embeddings, fixed-length vectors extracted from a pretrained protein language model (ESM1b) (Rives et al., 2021). We show that pretrained protein embeddings are informative for mAbs property prediction. Supervised learning using simple machine learning models and small labeled dataset suffice to enrich the signals and learn the sequence-to-solubility relationship. We also find that mAb solubility behavior differs among different IgG subclasses. solPredict can predict IgG1 and IgG4 reliably, but more IgG2 data are needed to be generalizable to IgG2. In this work, we provide a systematic framework for the study of antibody developability when limited data are available, which can be used for high-throughput screening of mAbs during early antibody discovery.

RESULTS

Dataset construction

We sought to construct a large, diverse dataset with quantitative measures of solubility behavior. We used polyethylene glycol 3350 (PEG 3350) with concentration ranging from 4% to 36% to assess the relative solubility of 260 mAbs (see STAR Methods for details). We chose the PEG-induced approach due to its capability of robust, high-throughput screening with minimal material consumption (Chai et al., 2019). Out of the 260 mAbs, 112 are IgG1, 140 are IgG4, and 8 are IgG2 molecules. Measurements were made in a 0.4 M L-histidine (pH 6.0) buffer condition, which has emerged as a popular choice for mAb-based products (Kingsbury et al., 2020). To investigate the intrinsic experimental noise, we examined two control mAbs with known solubility behaviors (mAb239: IgG1, good solubility; mAb240: IgG4, poor solubility) (Table 1). The two control mAbs were measured under histidine buffer pH 6.0 (H6) condition for 47 times. The summary statistics are shown in Table 1 and the distribution of solubility measurements is reported in Figure 1A. The measurement of the well-behaved control (mAb239) is 35.81 ± 0.72 PEG % while the measurement of the poorly behaved control (mAb240) is 10.63 ± 1.46 PEG %, demonstrating the robustness of the PEG-induced approach to differentiate mAbs by their solubility.

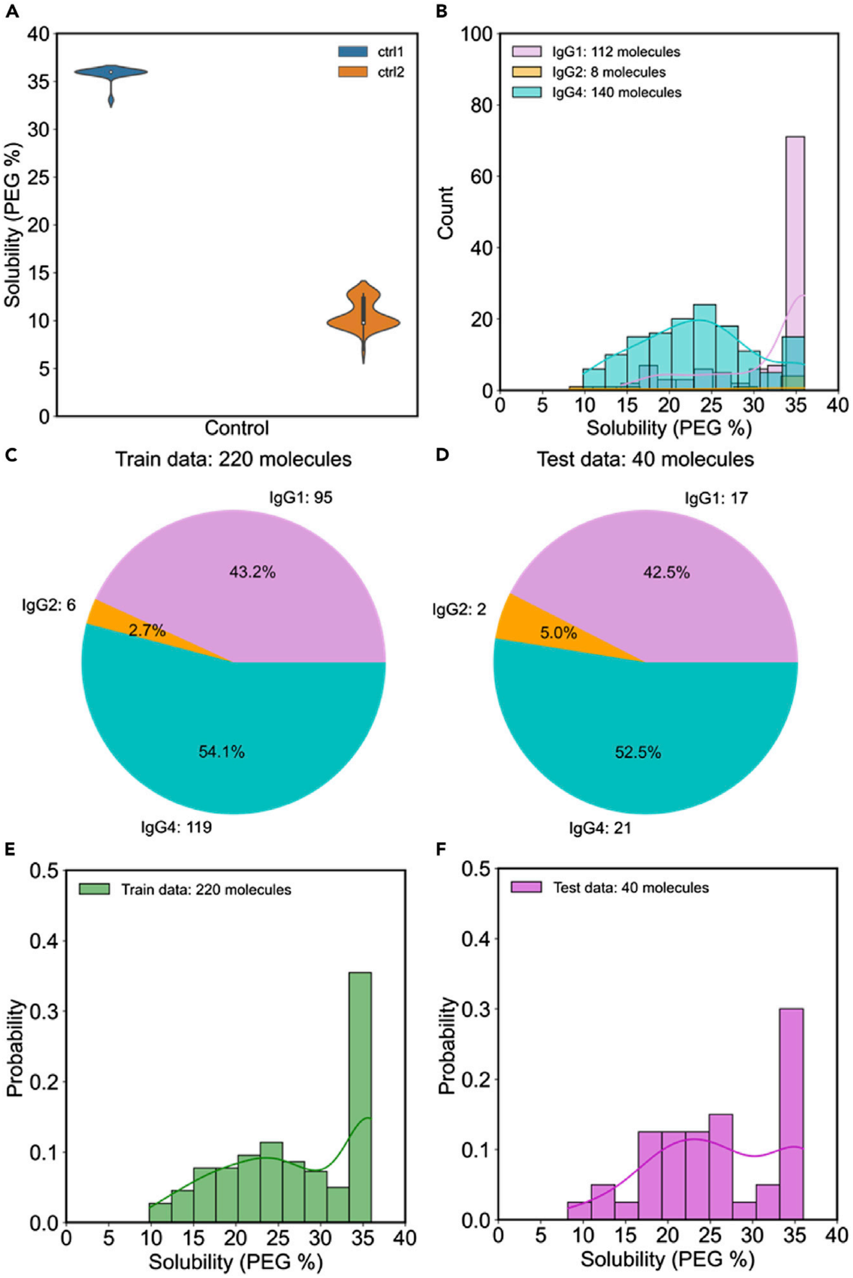


Figure 1. mAb solubility dataset

(A) The solubility distribution of two controls at histidine buffer pH 6.0 (H6). For each control, the solubility measurements were repeated for 47 times.

(B) The solubility distribution of different IgG subclasses. The IgG subclass composition for (C) training dataset (n = 220) and (D) test dataset (n = 40). The solubility distribution for (E) training dataset (n = 220) and (F) test dataset (n = 40). See also Figures S1 and S2.

Next, we explored the contribution of antibody subclass to solubility behavior (Figures 1B and S1). Previous developability studies focused on the variable domains due to the sequence and structural homology between IgG subclasses (Jain et al., 2017; Raybould et al., 2019). However, our results suggest IgG1 and IgG4 exhibit divergent solubility behavior. A broad range of solubility was observed in IgG4 molecules (median of 140 molecules is 23.6 PEG %), whereas IgG1 mAbs tend to be highly soluble (median of 112 molecules is 36.0 PEG %) under H6 condition. Similarly, previous study reports that IgG1 and IgG4 mAbs show totally

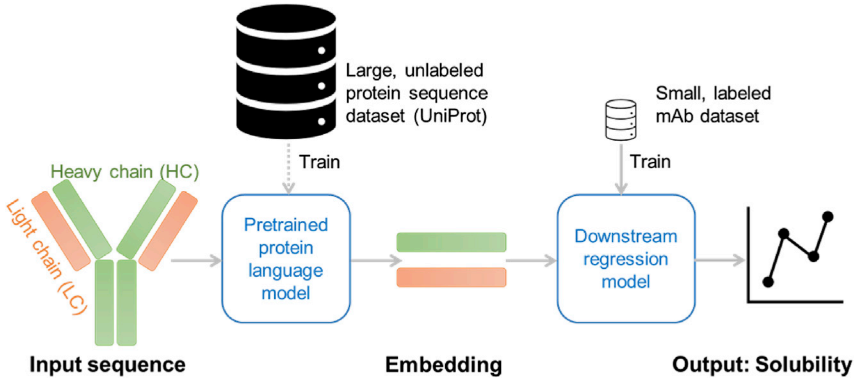

Large, unlabeled protein sequence dataset (UniProt)
Small, labeled mAb dataset
Heavy chain (HC)
Light chain (LC)
Train
Train
Pretrained protein language model
Downstream regression model
Input sequence
Embedding
Output: Solubility

Figure 2. The solPredict architecture

First, full IgG sequences were converted into fixed size embeddings (1280 for each chain) through protein language model pretrained using large, unlabeled protein sequence database. Next, the heavy chain and light chain embeddings were concatenated into 2560-dimensional feature vectors and used as the input of a downstream regression model that predicts mAb solubility. Quantitative solubility data measured by PEG-induced precipitation method were used to supervise the training of the regression model.

different behavior in terms of viscosity, opalescence, thermostability, and solubility (Bailly et al., 2020; Han et al., 2022; Kingsbury et al., 2020; Starr et al., 2021). To identify the underlying molecular determinants of antibody solution behavior, Kingsbury et al. compiled a diverse dataset of 59 mAbs and 23 molecular descriptors spanning colloidal, conformational, charge-based, hydrodynamic, and hydrophobic properties (Kingsbury et al., 2020). The authors reported similar distribution of many characteristics among IgG1s, IgG2s, and IgG4s (i.e. solution charge, hydrophobicity indices, and Fv-CSP). The major difference is the charge at the constant regions. The calculated charge of the IgG4 Fc dimer at pH 6.0 is negative (−1.92), whereas the constant regions of IgG1s (+6.50) and IgG2s (+5.22) are positively charged. The Fv region at pH 6.0 is generally positively charged. It is possible that the repulsion of positive charges between Fv and Fc leads to the higher solubility of IgG1s compared with IgG4s. Additionally, the authors suggest that mAbs with higher isoelectric points (pI) values exhibit better solution behavior. A strong positive correlation is also observed between experimental solubility and pI (Pearson's r = 0.8 for all IgGs, Pearson's r = 0.7 for IgG1s, Pearson's r = 0.6 for IgG4s) (Figure S2). Generally, mAbs with higher pI exhibit better solubility behavior.

Therefore, a dataset consisting of different subclasses is critical to the development of computational predictors that can be generalizable to diverse mAbs. To ensure the similar subclass composition of train/test dataset, we randomly split the dataset into train and test group (85/15) for each subclass (Figures 1C–1F). The train dataset contains 220 mAbs, consisting of 95 IgG1, 6 IgG2, and 119 IgG4 (Figure 1C). The test dataset contains 40 mAbs, composed of 17 IgG1, 2 IgG2, and 21 IgG4 (Figure 1D). The final solubility distribution is similar between train and test set (Figures 1E and 1F).

Pretrained protein language model embeddings enable the use of small, labeled dataset and simple architecture for mAbs solubility prediction

While the current dataset of 260 mAbs is already larger than previous mAbs developability studies (Bailly et al., 2020; Coffman et al., 2020; Han et al., 2022; Jain et al., 2017; Kingsbury et al., 2020; Lai et al., 2021a, 2021b; Raybould et al., 2019; Zhang et al., 2020), it is still smaller than most machine learning projects in which millions of data points were used for training (Altae-Tran et al., 2017). To overcome the data scarcity, we propose an end-to-end machine learning framework for solubility prediction with transfer learning from pretrained protein language model (ESM1b) (Rives et al., 2021) (Figure 2). ESM1b model was trained on 86 billion amino acids across 250 million unlabeled protein sequences in an unsupervised manner. The learned representations contain rich information about biological properties, ranging from biochemical properties of amino acids to the secondary and tertiary structure. Our proposed framework is mainly empowered by the informative protein representation from ESM1b, which captures the general sequence semantics in the protein universe. At first, we extracted embeddings for both heavy and light chains from the last layer of ESM1b. Second, we used the concatenated embeddings (2560-dimensional feature vector, 1280 for each chain) as input and trained various downstream regression models for solubility prediction.

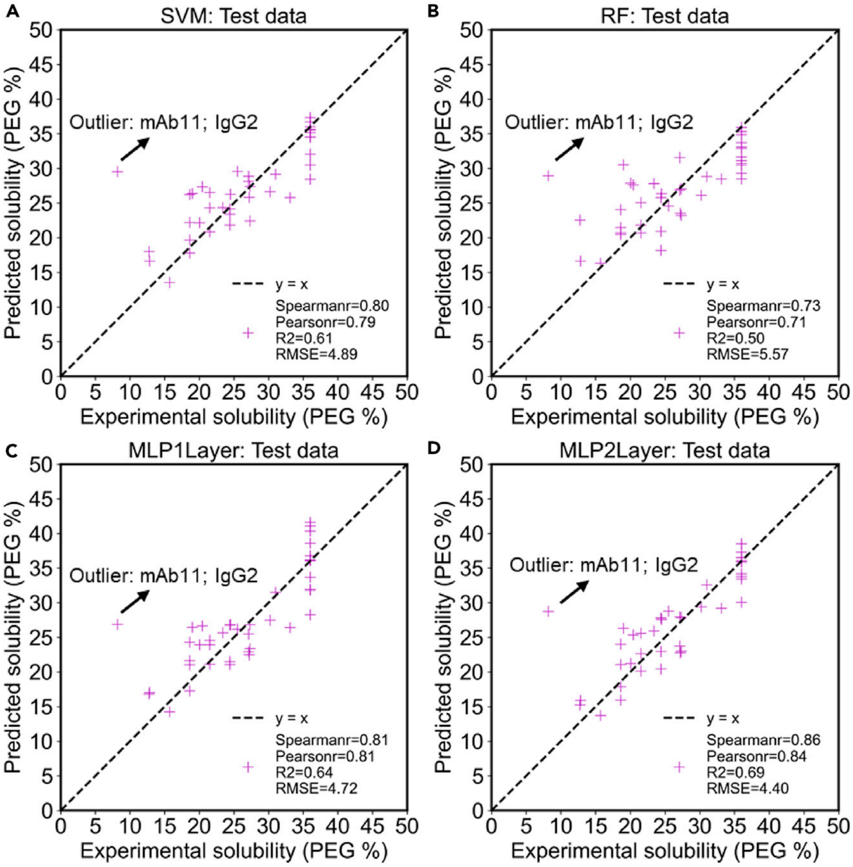


Figure 3. Performance of different regression models on test dataset (n = 40)

The correlation between the predicted (y axis) and experimentally measured solubility (x axis) on the test dataset for (A) SVM, (B) RF, (C) MLP1Layer, and (D) MLP2Layer models. The dashed black line refers to the perfect correlation: y = x. The statistics of four evaluation metrics are shown in legend. The outlier (mAb11, IgG2) is annotated for each model. See also Table S2, Figures S4–S7.

We compared traditional machine learning models (support vector machine regressor (SVM), random forest regressor (RF)) to neutral network models (multilayer perceptron (MLP) with 1 fully connected hidden layer (MLP1Layer), MLP with 2 fully connected hidden layers (MLP2Layer)). For SVM and RF, the input embeddings were first compressed into 23-dimensional vectors using principal component analysis (PCA) to explain >90% variance. For MLP1Layer and MLP2Layer, the input embeddings were directly used as the first layer of the network. 5-fold cross validation of the training dataset of 220 mAbs was used for hyperparameter tuning for all models. Once the optimal hyperparameter set was determined, the final models for SVM and RF were trained on the whole training set using these hyperparameters. For MLP1Layer and MLP2Layer model, Spearman correlation coefficient on validation set was used to select the best checkpoint during training. The average Spearman correlation coefficient was used to select the optimal combination of hyperparameter and the resulting five models. Instead of refitting with the whole training set, the mean of five models was used for prediction. The hyperparameter search range and the selected combination of hyperparameter are shown in Table S1 and Figure S3. Full one-hot protein encoding (14,343-dimensional vector per mAb which is around 5.6-fold larger than ESM1b embedding)-based SVM and MLP2Layer models were used as two baseline models.

The test dataset of 40 mAbs was used to compare these six models (Table S2, Figures 3, S4, and S5). We reported a variety of performance metrics for systematic comparisons: Spearman correlation coefficient, Pearson correlation coefficient, R^2, and root-mean-square error (RMSE). As shown in Table S2, ESM1b embedding improves overall performance for both SVM and MLP2Layer models. The improvement is most significant for SVM model, with Spearman correlation coefficient increases from 0.60 to 0.80, Pearson correlation coefficient increases from 0.62 to 0.79, R^2 increases from 0.38 to 0.61, and RMSE reduces from 6.20 to 4.89. For the MLP2Layer model, ESM1b embedding improves Spearman correlation coefficient from 0.82 to 0.86, Pearson correlation coefficient from 0.82 to 0.84, R^2 from 0.66 to 0.69, and reduces RMSE from 4.60 to 4.40. It is important to note that the simple ESM1b-based SVM model is comparable with full one-hot protein encoding-based MLP2Layer model (Spearman correlation coefficient: 0.80 versus 0.82, Pearson correlation coefficient: 0.79 versus 0.82, R^2: 0.61 versus 0.66, and RMSE: 4.89 versus 4.60) (Table S2). This suggests that pretrained embeddings are effectively informative that simple machine learning models supervised with a small annotated experimental dataset suffice to predict antibody solubility with high performance. As MLP2Layer model performs the best based on all four evaluation metrics, it was selected as the downstream regression model for solPredict.

To evaluate whether the high performance of solPredict (ESM1b-based MLP2Layer model) is due to chance, Y scrambling was performed for five times where training data labels were shuffled to create fake feature-label pairs (Figure S6). The original average Spearman correlation coefficient during training is 0.86 (Figure S3). However, the average Spearman correlation coefficients for all five Y scrambling experiments drop to ~0.2 (Figure S6A). No statistically significant correlations were observed between the new predictions and experimental solubility on the 40 test molecules (Figures S6B–S6F). Overall, this suggests that the high performance is due to the learned sequence-to-solubility relationship instead of pure chance.

To further investigate the effect of training sample size on the performance of solPredict (ESM1b-based MLP2Layer model), we varied the training sample size ratio from 0.1 to 1 of the original 220 training molecules and evaluated the model performance on the 40 test molecules (Figure S7). With increasing training sample size, all four evaluation metrics improve and gradually converge around training sample size ratio = 0.7 (154 training molecules). This suggests that a smaller training sample size may be sufficient when using a protein language model. Taken together, we show that protein language model pretrained on general protein sequences can provide a powerful signal for antibody-related downstream task with limited labeled dataset.

solPredict can predict both IgG1 and IgG4 accurately

Next, we investigated how performance varies among different isotypes (Figure 4). Our analysis shows that solPredict can predict both IgG1 (n = 17) and IgG4 (n = 21) with high performance (Figures 4A, 4B, and Table S3). The performance on the IgG1 test set is the best with Spearman correlation coefficient = 0.84, Pearson correlation coefficient = 0.93, R^2 = 0.87, and RMSE = 2.36. The performance on the IgG4 test set is slightly worse than IgG1 with Spearman correlation coefficient = 0.69, Pearson correlation coefficient = 0.77, R^2 = 0.58, and RMSE = 3.48. This is due to the different solubility behavior between IgG1 and IgG4 molecules. Most IgG1s exhibit high solubility (>30 PEG %), while IgG4s exhibit a bell curve distribution, with the bulk showing medium solubility (~25 PEG %) (Figures 1B, 4A, 4B, and S1). The relationship between sequence and solubility is therefore simpler to learn for IgG1 compared with IgG4 molecules.

For IgG2 (n = 2), there is an outlier, mAb11, whose solubility was overestimated by all four models (Figures 3A–3D and 4C). The experimental solubility for this IgG2 molecule is 8.20 PEG %, whereas all four models consistently predicted it around 30 PEG %. We sought to understand why mAb11 cannot be reliably predicted. We first compared the embeddings of mAb11 against all 220 training molecules and selected the top 10 mAbs that are closest to mAb11 based on the Euclidean distance (Figure 4D and Table 2). 7 out of the 10 closest neighbors are IgG4 and the closest neighbor is mAb5, an IgG4 molecule with high solubility (31.80 PEG %). Only 1 out of the 10 closest neighbors is IgG2, which also shows high solubility (36.00 PEG %). It appears that solPredict predicts mAb11 to be much more soluble than the experimental measurement because it reasons with the mapping relationship learned from IgG4 molecules. It suggests that the sequence-to-solubility mapping relationship for IgG2 is different from IgG1 and IgG4, and solPredict could not learn the sequence-to-solubility mapping for IgG2 due to the limited IgG2 molecules (n = 6) in the training set. To sum up, solPredict can serve as an accurate predictor of IgG1 and IgG4 mAb solubility, while more IgG2 data are needed to ensure the generalizability to IgG2 prediction.

Pretrained protein embeddings are informative for antibody solubility behavior

To interrogate how solPredict learns to predict mAb solubility, we projected the embeddings into two dimensions with t-distributed stochastic neighbor embedding (t-SNE) (Maaten and Hinton, 2008). We performed the analysis both for the raw embeddings from ESM1b (Figure 5A) and for the final hidden layer of the MLP2Layer

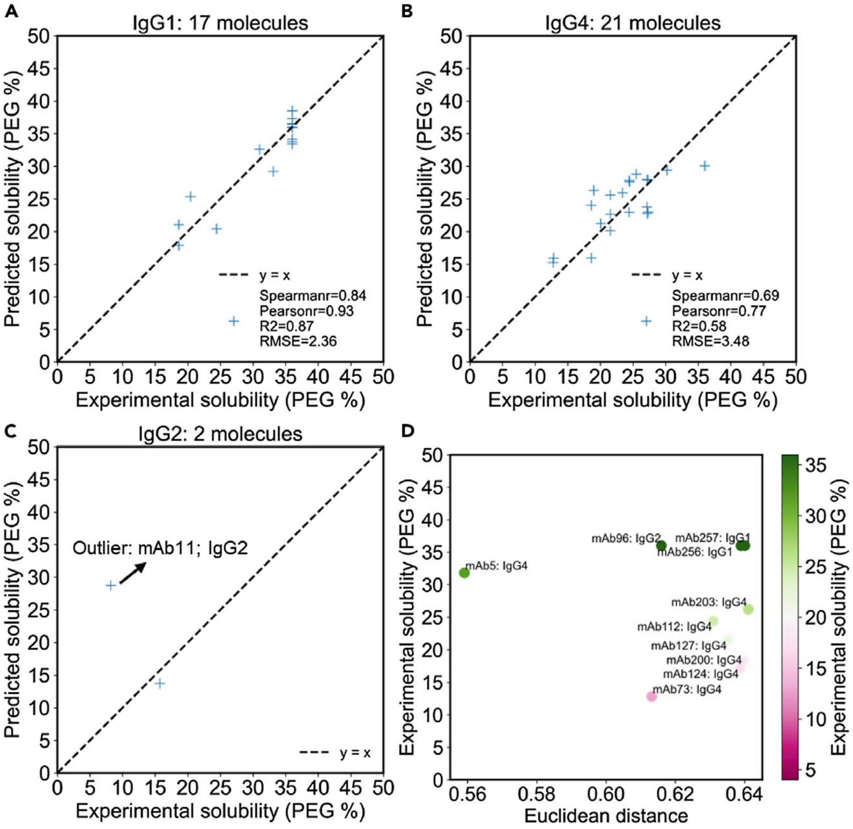


Figure 4. Performance of solPredict (MLP2Layer as downstream regressor) on different IgG subclasses

The correlation between the predicted (y axis) and experimentally measured solubility (x axis) for (A) IgG1 test dataset (n = 17), (B) IgG4 test dataset (n = 21), and (C) IgG2 test dataset (n = 2). The dashed black line represents the perfect correlation: y = x. The values of four evaluation metrics are shown in legend for IgG1 and IgG4. Due to the limited sample size of IgG2 (n = 2), no evaluation metric was computed. The outlier (mAb11, IgG2) is highlighted in (C).

(D) Visualization of 10 training molecules closest to mAb11. The x axis represents the Euclidean distance of embeddings between mAb11 and the other mAbs. The y axis represents the experimentally determined solubility. Each data point is colored based on the experimentally determined solubility and annotated by the mAb index and subclass. See also Table S3.

network after training (Figure 5B). The key hyperparameter (perplexity) was optimized for t-SNE construction (see STAR Methods for details). All t-SNE representations in Figure 5 were created using 5000 iterations, PCA as initialization, and Euclidean distance as similarity metric. The perplexity was set as 10 for the raw representation and 30 for the last hidden layer representation. As a baseline comparison, the same methodology and hyperparameter were used to construct two-dimensional t-SNE representations of one-hot encoding (perplexity = 10) and the last layer (perplexity = 30) of one-hot encoding-based MLP2Layer model (Figures S5E and S5F). First, the t-SNE representations were colored according to their experimental solubility at H6 (Figures 5, S5E, and S5F). Although never trained, the raw embeddings appeared to capture some information about mAb solubility, with diffuse organization of small communities sharing similar solubility behavior (Figure 5A). After training, the network learned the sequence-to-solubility relationship and showed clear structure of representation space with the diagonal corresponding to the variation in solubility (Figure 5B). The outlier mAb11 was incorrectly placed in the middle of high-solubility mAbs, further verifying our previous hypothesis that the relationship of IgG2 was not well learned yet. Similar patterns were observed for raw one-hot encoding and trained one-hot encoding-based MLP2Layer model (Figures S5E and S5F). This may suggest that mAbs with similar sequences tend to exhibit similar solubility behavior and neutral network can learn the complex sequence-to-solubility relationship.

To evaluate whether ESM1b embeddings learned a biologically meaningful representation of antibody sequences, the t-SNE representations of raw embeddings and one-hot encoding were annotated based on

Table 2. Top 10 mAbs in the train dataset closest to mAb11 ranked by the Euclidean distance

mAb	Euclidean distance	Subclass	H6 (PEG %)
mAb11	0	IgG2	8.20
mAb5	**0.56**	**IgG4**	**31.80**
mAb73	0.61	IgG4	12.80
mAb96	**0.62**	**IgG2**	**36.00**
mAb112	0.63	IgG4	24.40
mAb127	0.64	IgG4	21.50
mAb124	0.64	IgG4	17.15
mAb256	0.64	IgG1	36.00
mAb200	0.64	IgG4	18.30
mAb257	0.64	IgG1	36.00
mAb203	0.64	IgG4	26.20

heavy chain and light chain V gene segments (Figure S8). In contrast to the one-hot encoding (Figures S8C and S8D), ESM1b embeddings naturally separate according to the underlying V gene segments (Figures S8A and S8B). This suggests that ESM1b embeddings captured functionally important information about antibodies and ESM1b embeddings can serve as a better representation of antibodies compared with one-hot encoding. This may explain why the simple ESM1b-based SVM model is comparable with one-hot encoding-based neutral network model (MLP2Layer). Together, it shows that pretrained protein embeddings contain meaningful representations of antibody sequences and further supervised learning with a small dataset suffice to reorganize the embeddings for specific task.

DISCUSSION

We propose that incorporating high-capacity protein language model pretrained on 100s of millions of sequences stored in protein databases (e.g. UniProt (The UniProt Consortium, 2021)) will be the key to alleviate both the data scarcity and feature engineering challenges for antibody developability prediction. As a proof-of-concept, we applied the pretrained protein language model to predict mAb solubility through transfer learning. We represented full IgG sequences using embeddings extracted from the pretrained protein language model. Using the embeddings as input, we trained four simple machine learning models (SVM, RF, MLP1Layer, and MLP2Layer) with a diverse set of 220 mAbs. Methods were compared using an independent test set of 40 mAbs. We find that all four models show high correlation with experimental measurements despite the simplicity of models. MLP2Layer model performs the best with Spearman correlation coefficient = 0.86, Pearson correlation coefficient = 0.84, $R^2 = 0.69$, and RMSE = 4.40. Furthermore, MLP2Layer model can predict both IgG1 and IgG4 mAbs with high performance despite their distinct solubility behavior. Our results suggest that pretrained protein embeddings are powerful representations for sequence-to-property mapping for mAbs. Further supervised learning with small, labeled dataset can enhance the signal for specific task, solubility prediction in this case. We anticipate that transfer learning and massive protein language models can also be used to predict other developability properties, such as viscosity, aggregation propensity, and stability.

solPredict can reliably predict antibody apparent solubility from sequence information alone, which makes it suitable for screening the solubility of a large library of mAbs during early antibody discovery. In general, PEG $\leq$ 19% is used as guideline for flagging mAbs with poor solubility in the screening assay. With 19 PEG % as threshold, solPredict achieves a successful classification rate of ~90% on the test dataset (Figure S9A). The errors mainly come from false positives. 4 out of 40 mAbs were misclassified as high soluble (1 IgG1, 2 IgG4, and 1 IgG2 molecules) while none of them was misclassified as low soluble (Figure S9). It suggests that solPredict will be less likely to eliminate otherwise highly soluble mAb candidates, which is more costly and consequential than false positive error.

In summary, our work confirms the value of pretrained protein language models on antibody properties and demonstrates a new framework for rapid and high-throughput antibody developability prediction using only sequence information. We anticipate that this new framework will facilitate the early developability screening of antibodies and minimize the cost and time needed for the selection of lead mAb candidates.

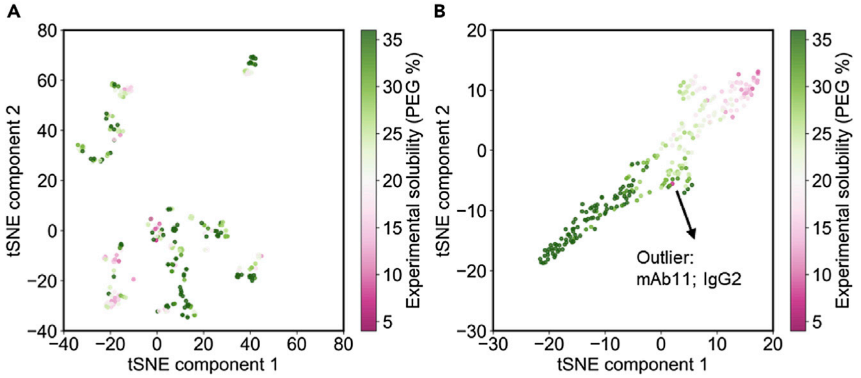


Figure 5. Visualization of embeddings along two dimensions using t-SNE

The representation of all 260 mAbs using (A) raw embeddings and (B) last hidden layer of MLP2Layer model after training. The outlier (mAb11, IgG2) is highlighted in (B). Each point represents a mAb and each mAb is colored by the experimentally measured solubility. See also Figures S5 and S8.

Limitations of the study

There are three major limitations of this study. The first limitation is the generalizability to new molecules which are very different from the training dataset. For example, the current model overestimates the solubility of one IgG2 molecule due to the limited IgG2 molecules (n = 6) in the training set (Figure S4). A potential solution is constantly updating the solubility model with the continuous incoming streams of data. Another limitation of this study is the interpretability of the model. The extracted embeddings from pre-trained protein language model capture multiscale properties of the protein, ranging from biochemical properties of amino acids to remote homology of proteins. Therefore, it is challenging to offer mechanistic insights into the mAb solubility behavior. Future work involving descriptor analysis will be needed to understand the key contributors to poor solubility behavior. Finally, current model is not sensitive to predict the impact of single mutants on solubility (Figure S10). Shan et al. reported the apparent solubility of 16 mAb-J variants under similar buffer condition (25 mM histidine/histidine HCl, 5 mM arginine HCl, pH 6.0 buffer) (Shan et al., 2018). The full-length sequences of 16 molecules were fed into the solPredict model and the predictions were compared against the reported experimental value (Figure S10). However, no statistically significant correlation was observed between predicted and experimental results. This may be because only protein level embedding (average of all amino acid features) was used in this work, which will cause the loss of amino acid level information.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Antibodies
  - ○ PEG-induced precipitation for solubility measurement
  - ○ Protein embeddings computation
  - ○ Training details
  - ○ Baselines
  - ○ Y scrambling
  - ○ Performance evaluations
  - ○ t-SNE construction
  - ○ Computation of antibody isoelectric point (pI)

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.105173.

ACKNOWLEDGMENTS

We thank Peter Hillier and Yuhao Lin for valuable software support. We thank Guilherme Rocha, William Smith, and Ahmadreza Ghanbarpour Ghouchani for helpful discussion. We thank the antibody engineering group and automation group at Lilly Biotechnology Center for characterizing, producing, and purifying the antibodies.

AUTHOR CONTRIBUTIONS

Q.C. and J.F. conceived the idea. Q.C. supervised the research. J.S. conducted the PEG-induced precipitation experiment for solubility measurement. M.J. built the ESM1b environment and contributed to the design of transfer learning strategies. J.F. analyzed the data, established the machine learning models, and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 3, 2022
Revised: August 16, 2022
Accepted: September 14, 2022
Published: October 21, 2022

REFERENCES

Altae-Tran, H., Ramsundar, B., Pappu, A.S., and Pande, V. (2017). Low data drug discovery with one-shot learning. ACS Cent. Sci. *3*, 283–293. https://doi.org/10.1021/acscentsci.6b00367.

Anselmo, A.C., Gokarn, Y., and Mitragotri, S. (2019). Non-invasive delivery strategies for biologics. Nat. Rev. Drug Discov. *18*, 19–40. https://doi.org/10.1038/nrd.2018.183.

Bailly, M., Mieczkowski, C., Juan, V., Metwally, E., Tomazela, D., Baker, J., Uchida, M., Kofman, E., Raoufi, F., Motlagh, S., et al. (2020). Predicting antibody developability profiles through early stage discovery screening. mAbs *12*, 1743053. https://doi.org/10.1080/19420862.2020.1743053.

Bepler, T., and Berger, B. (2021). Learning the protein language: evolution, structure, and function. Cell Syst. *12*, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

Chai, Q., Shih, J., Weldon, C., Phan, S., and Jones, B.E. (2019). Development of a high-throughput solubility screening assay for use in antibody discovery. mAbs *11*, 747–756. https://doi.org/10.1080/19420862.2019.1589851.

Chan, A.C., and Carter, P.J. (2010). Therapeutic antibodies for autoimmunity and inflammation. Nat. Rev. Immunol. *10*, 301–316. https://doi.org/10.1038/nri2761.

Chan, P., Curtis, R.A., and Warwicker, J. (2013). Soluble expression of proteins correlates with a lack of positively-charged surface. Sci. Rep. *3*, 3333. https://doi.org/10.1038/srep03333.

Chemical Computing Group (2022). Molecular Operating Environment (MOE).

Coffman, J., Marques, B., Orozco, R., Aswath, M., Mohammad, H., Zimmermann, E., Khouri, J., Griesbach, J., Izadi, S., Williams, A., et al. (2020). Highland games: a benchmarking exercise in predicting biophysical and drug properties of monoclonal antibodies from amino acid sequences. Biotechnol. Bioeng. *117*, 2100–2115. https://doi.org/10.1002/bit.27349.

Dean, A.Q., Luo, S., Twomey, J.D., and Zhang, B. (2021). Targeting cancer with antibody-drug conjugates: promises and challenges. mAbs *13*, 1951427. https://doi.org/10.1080/19420862.2021.1951427.

Han, X., Shih, J., Lin, Y., Chai, Q., and Cramer, S.M. (2022). Development of QSAR models for in silico screening of antibody solubility. mAbs *14*, 2062807. https://doi.org/10.1080/19420862.2022.2062807.

Hebditch, M., Carballo-Amador, M.A., Charonis, S., Curtis, R., and Warwicker, J. (2017). Protein–Sol: a web tool for predicting protein solubility from sequence. Bioinformatics *33*, 3098–3100. https://doi.org/10.1093/bioinformatics/btx345.

Hunter, J.D. (2007). Matplotlib: a 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Jain, T., Sun, T., Durand, S., Hall, A., Houston, N.R., Nett, J.H., Sharkey, B., Bobrowicz, B., Caffry, I., Yu, Y., et al. (2017). Biophysical properties of the clinical-stage antibody landscape. Proc. Natl. Acad. Sci. USA *114*, 944–949. https://doi.org/10.1073/pnas.1616408114.

Kingsbury, J.S., Saini, A., Auclair, S.M., Fu, L., Lantz, M.M., Halloran, K.T., Calero-Rubio, C., Schwenger, W., Airiau, C.Y., Zhang, J., et al. (2020). A single molecular descriptor to predict solution behavior of therapeutic antibodies. Sci. Adv. *6*, eabb0372. https://doi.org/10.1126/sciadv.abb0372.

Lai, P.-K., Fernando, A., Cloutier, T.K., Kingsbury, J.S., Gokarn, Y., Halloran, K.T., Calero-Rubio, C., and Trout, B.L. (2021a). Machine learning feature selection for predicting high concentration therapeutic antibody aggregation. J. Pharm. Sci. *110*, 1583–1591. https://doi.org/10.1016/j.xphs.2020.12.014.

Lai, P.-K., Fernando, A., Cloutier, T.K., Gokarn, Y., Zhang, J., Schwenger, W., Chari, R., Calero-Rubio, C., and Trout, B.L. (2021b). Machine learning applied to determine the molecular descriptors responsible for the viscosity behavior of concentrated therapeutic antibodies. Mol. Pharm. *18*, 1167–1175. https://doi.org/10.1021/acs.molpharmaceut.0c01073.

Leavy, O. (2010). Therapeutic antibodies: past, present and future. Nat. Rev. Immunol. *10*, 297. https://doi.org/10.1038/nri2763.

Li, H., Robertson, A.D., and Jensen, J.H. (2005). Very fast empirical prediction and rationalization of protein pKa values. Proteins *61*, 704–721. https://doi.org/10.1002/prot.20660.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Makowski, E.K., Wu, L., Gupta, P., and Tessier, P.M. (2021). Discovery-stage identification of drug-like antibodies using emerging experimental and computational methods. mAbs *13*, 1895540. https://doi.org/10.1080/19420862.2021.1895540.

Meireles Ribeiro, J., and Sillero, A. (1991). A program to calculate the isoelectric point of macromolecules. Comput. Biol. Med. *21*, 131–141. https://doi.org/10.1016/0010-4825(91)90022-2.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.

(2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., and Song, Y. (2019). Evaluating protein transfer learning with TAPE. In Advances in Neural Information Processing Systems (Curran Associates, Inc.).

Raybould, M.I.J., Marks, C., Krawczyk, K., Taddese, B., Nowak, J., Lewis, A.P., Bujotzek, A., Shi, J., and Deane, C.M. (2019). Five computational developability guidelines for therapeutic antibody profiling. Proc. Natl. Acad. Sci. USA *116*, 4025–4030. https://doi.org/10.1073/pnas.1810576116.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc. Natl. Acad. Sci. USA *118*. e2016239118. https://doi.org/10.1073/pnas.2016239118.

Shan, L., Mody, N., Sormani, P., Rosenthal, K.L., Damschroder, M.M., and Esfandiary, R. (2018). Developability assessment of engineered monoclonal antibody variants with a complex self-association behavior using complementary analytical and in silico tools. Mol. Pharm. *15*, 5697–5710. https://doi.org/10.1021/acs.molpharmaceut.8b00867.

Sharma, V.K., Patapoff, T.W., Kabakoff, B., Pai, S., Hilario, E., Zhang, B., Li, C., Borisov, O., Kelley, R.F., Chorny, I., et al. (2014). In silico selection of therapeutic antibodies for development: viscosity, clearance, and chemical stability. Proc. Natl. Acad. Sci. USA *111*, 18601–18606. https://doi.org/10.1073/pnas.1421779112.

Shire, S.J., Shahrokh, Z., and Liu, J. (2004). Challenges in the development of high protein concentration formulations. J. Pharm. Sci. *93*, 1390–1402. https://doi.org/10.1002/jps.20079.

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. (2012). PROSO II – a new method for protein solubility prediction. FEBS J. *279*, 2192–2200. https://doi.org/10.1111/j.1742-4658.2012.08603.x.

Sormanni, P., Aprile, F.A., and Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. J. Mol. Biol. *427*, 478–490. https://doi.org/10.1016/j.jmb.2014.09.026.

Sormanni, P., Amery, L., Ekizoglou, S., Vendruscolo, M., and Popovic, B. (2017). Rapid and accurate in silico solubility screening of a monoclonal antibody library. Sci. Rep. *7*, 8200. https://doi.org/10.1038/s41598-017-07800-w.

Starr, C.G., Makowski, E.K., Wu, L., Berg, B., Kingsbury, J.S., Gokarn, Y.R., and Tessier, P.M. (2021). Ultradilute measurements of self-association for the identification of antibodies with favorable high-concentration solution properties. Mol. Pharm. *18*, 2744–2753. https://doi.org/10.1021/acs.molpharmaceut.1c00280.

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489. https://doi.org/10.1093/nar/gkaa1100.

Thorsteinson, N., Gunn, J.R., Kelly, K., Long, W., and Labute, P. (2021). Structure-based charge calculations for predicting isoelectric point, viscosity, clearance, and profiling antibody therapeutics. mAbs *13*, 1981805. https://doi.org/10.1080/19420862.2021.1981805.

Trainor, K., Broom, A., and Meiering, E.M. (2017). Exploring the relationships between protein sequence, structure and solubility. Curr. Opin. Struct. Biol. *42*, 136–146. https://doi.org/10.1016/j.sbi.2017.01.004.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Waskom, M. (2021). seaborn: statistical data visualization. J. Open Source Softw. *6*, 3021. https://doi.org/10.21105/joss.03021.

Weiner, L.M., Surana, R., and Wang, S. (2010). Monoclonal antibodies: versatile platforms for cancer immunotherapy. Nat. Rev. Immunol. *10*, 317–327. https://doi.org/10.1038/nri2744.

Wolf Pérez, A.M., Sormanni, P., Andersen, J.S., Sakhnini, L.I., Rodriguez-Leon, I., Bjelke, J.R., Gajhede, A.J., De Maria, L., Otzen, D.E., Vendruscolo, M., et al. (2019). In vitro and in silico assessment of the developability of a designed monoclonal antibody library. mAbs *11*, 388–400. https://doi.org/10.1080/19420862.2018.1556082.

Wolf Pérez, A.-M., Lorenzen, N., Vendruscolo, M., and Sormanni, P. (2022). Assessment of therapeutic AntibodyTherapeutic antibodies DevelopabilityDevelopability by combinations of in vitro and in SilicoIn silico methods. In Therapeutic Antibodies: Methods and Protocols, G. Houen, ed. (Springer US), pp. 57–113.

Zhang, Y., Wu, L., Gupta, P., Desai, A.A., Smith, M.D., Rabia, L.A., Ludwig, S.D., and Tessier, P.M. (2020). Physicochemical rules for identifying monoclonal antibodies with drug-like specificity. Mol. Pharm. *17*, 2555–2569. https://doi.org/10.1021/acs.molpharmaceut.0c00257.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python version 3.8	Python Software Foundation	https://www.python.org/
ESM1b	Rives et al., 2021	https://github.com/facebookresearch/esm
Scikit-learn	Pedregosa et al., 2011	https://scikit-learn.org/stable/
PyTorch	Paszke et al., 2019	https://pytorch.org/
SciPy	Virtanen et al., 2020	https://scipy.org/
t-distributed stochastic neighbor embedding (t-SNE)	Maaten and Hinton, 2008	https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html
Molecular Operating Environment (MOE)	Chemical Computing Group, 2022	https://www.chemcomp.com/index.htm
Matplotlib	Hunter, 2007	https://matplotlib.org/stable/index.html
seaborn	Waskom, 2021	https://seaborn.pydata.org/index.html
Model codes	Github	https://github.com/JiangyanFeng-Lilly/solPredict_manuscript_codes_2022

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Qing Chai (chai_qing_qc@lilly.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The antibody sequence data used in this study cannot be deposited in a public repository because they are confidential information.
- All original code has been deposited at https://github.com/JiangyanFeng-Lilly/solPredict_manuscript_codes_2022 and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Antibodies

A total of 260 in-house mAbs, consisting of 112 IgG1, 140 IgG4, and 8 IgG2 subclass were used in this study. They were all produced internally at Eli Lilly & Co using either HEK293 or Chinese hamster ovary cells expression system. mAbs were purified using a standard antibody purification procedure (protein A capture followed by polishing steps). All reagents and excipients were commercially available from Hampton Research, EM Chemicals, JT Baker, Sigma-Aldrich, and Mallinckrodt, and were of high purity (>98%).

PEG-induced precipitation for solubility measurement

The apparent solubilities of mAbs were measured using PEG-induced precipitation assays, as described in a previous publication (Chai et al., 2019). Stock solutions of 0.4 M L-histidine, pH 6.0 were used for H6 buffer matrices. Solutions with Polyethylene glycol 3350 (PEG 3350) levels varying from 4 to 36% were prepared and were mixed overnight at 1200 RPM at 25°C. All mAbs were buffer-exchanged and diluted with water to a target final concentration of 1 mg/mL. After mixing mAbs and stock solutions on each assay plate, all assay plates were sealed with foil and incubated at 25°C on the bench for 24 hours. Finally, the plates were read at 280 nm (with background subtraction at 320 nm) using a Tecan Infinite M1000 Pro UV/Vis

Spectrophotometer. The absorbance data was de-convoluted and plotted using Excel. The onset of precipitation was determined visually based on the point of abrupt decrease in absorbance caused by the loss of protein. The nearest PEG 3350 concentration (%) corresponding to the onset of precipitation was used as the estimation of solubility. Therefore, the experimental measurement of solubility varies from 4 to 36 PEG %. The higher the PEG %, the better is the solubility.

Protein embeddings computation

ESM1b model (Rives et al., 2021) was used as the pretrained protein language model to extract the embeddings for mAbs. ESM1b model was trained on 86 billion amino acids across 250 million protein sequences using unsupervised learning and improved a range of applications such as prediction of mutational effect, secondary structure, and long-range contacts. These representations therefore contain both sequence level and structural level signals. For each input sequence, ESM1b generates a 1280-dimensional vector representation for each amino acid. In this work, the average of all amino acid features was used as a final feature vector for each sequence (1280-dimensional vector per sequence). The full sequence of heavy chain and light chain for each IgG was used to generate 1280-dimensional vector per chain. The heavy chain and light chain embedding were then concatenated into a 2560-dimensional vector to represent each mAb.

Training details

For all four models, 220 mAbs were used for training and hyperparameter tuning through five-fold cross validation. For support vector machine and random forest regressor, we used PCA to reduce the dimensionality of the raw embeddings from 2560 to 23 to explain 90% of variance. Hyperparameters were optimized using a grid search and five-fold cross validation (Table S1). Once the hyperparameters were optimized, the whole training dataset of 220 mAbs was used to refit the support vector machine and random forest regressor. Scikit-learn (Pedregosa et al., 2011) was used to implement support vector machine and random forest models.

The neutral network models (MLP1Layer and MLP2Layer) were trained in a mini-batch mode using Pytorch (Paszke et al., 2019). MLP1Layer refers to multilayer perceptron with 1 fully connected hidden layer. MLP2Layer refers to multilayer perceptron with 2 fully connected hidden layers. ReLU was used as the activation function. The training process takes 500 epochs on the training dataset of 220 mAbs using the Adam optimizer and Xavier_uniform initialization. The hyperparameters (batch size, learning rate, and hidden layer dimensions) were tuned using five-fold cross validation (Table S1). The training objective is to minimize the mean squared error between predicted and experimental solubility values. To alleviate overfitting, the Spearman correlation coefficient on the validation set (the held-out fold) was used to select the best checkpoint during training. The average of Spearman correlation coefficient on validation sets was used to select the best combination of hyperparameters and the final set of models. The selected hyperparameters for MLP1Layer model are hidden layer size = 256, batch size = 32, and learning rate = 0.01. The selected hyperparameters for MLP2Layer model are hidden layer 1 size = 64, hidden layer 2 size = 32, batch size = 8, and learning rate = 0.001.

Baselines

Full one-hot protein encoding was used as the baseline featurization. For each IgG molecule, the full-length heavy and light chain sequence was first padded into target size (463 for heavy chain and 220 for light chain), then one-hot encoded, and finally concatenated into a vector of dimension (463+220) * 21 = 14343, where 21 represents 20 amino acid tokens plus one padding token. Compared with the ESM1b representation, the dimensionality of full one-hot protein encoding is around 5.6-fold larger. The same training methodology (as described in the Training details section) was used to construct the downstream SVM and MLP2Layer model. To eliminate the effect of hyperparameters, the same set of hyperparameters was used for one-hot encoding based MLP2Layer model: hidden layer 1 size = 64, hidden layer 2 size = 32, batch size = 8, and learning rate = 0.001.

Y scrambling

Y scrambling was used to evaluate whether the predictions made by the model are due to the chance. We first shuffled the labels of the training dataset to create the fake feature-label training dataset and then followed the same training methodology to train the MLP2Layer model using the fake feature-label pairs. The same set of hyperparameters was used, which is hidden layer 1 size = 64, hidden layer 2 size = 32, batch size = 8, and learning rate = 0.001. We conducted five rounds of Y scrambling and evaluated the performance of the new models on 40 test molecules.

Performance evaluations

The test set of 40 mAbs were never used for model training or hyperparameter tuning. The performance on test set were evaluated using Spearman correlation coefficient, Pearson correlation coefficient, R^2, and RMSE. Spearman and Pearson correlation coefficient were implemented using SciPy python package (Virtanen et al., 2020). Both correlation coefficients vary between −1 and +1 with 0 implying no correlation. The difference is that Pearson correlation assumes that data is normally distributed whereas Spearman correlation is a nonparametric measure without assuming that datasets are normally distributed. Considering the solubility data distribution is skewed in the dataset, Spearman correlation was used as the main evaluation metric. Scikit-learn (Pedregosa et al., 2011) was used to compute R^2 (sklearn.metrics.r2_score) and RMSE (square root of mean squared error: sklearn.metrics.mean_squared_error). R^2 (coefficient of determination) measures how well the regression predictions approximate the real data points with 1 indicating the perfect fit. R^2 can be negative if the model performs worse than a constant model (R^2 = 0). RMSE is a measure of the difference between predicted and actual values, with 0 indicating the perfect fit between predicted and actual data points. The lower the RMSD, the better the regression model. All figures in this work were generated using Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021).

t-SNE construction

To project the high-dimensional embeddings on to two-dimensional space, we used the t-distributed stochastic neighbor embedding (t-SNE) algorithm (Maaten and Hinton, 2008) implemented in Scikit-learn (Pedregosa et al., 2011). Perplexity parameter was varied from 5 to 100 with a step size of 5. Similarity metric was based on Euclidean distance. PCA was set as initialization and 5000 iterations were performed. In the end, the perplexity was selected as 10 for the raw representation and 30 for the last hidden layer representation.

Computation of antibody isoelectric point (pI)

Full-length homology models of all IgG molecules were generated and structure-based protein pI (Pro pI 3D) was computed using MOE software (Chemical Computing Group, 2022). The Pro pI 3D descriptor was calculated using PROPKA method to determine residue pKa values which are then used in the sequence-based pI formula (Li et al., 2005; Meireles Ribeiro and Sillero, 1991; Thorsteinson et al., 2021).