# SOLART: A STRUCTURE-BASED METHOD TO PREDICT PROTEIN SOLUBILITY AND AGGREGATION

Q. HOU [1], J. M. KWASIGROCH [1], M. ROOMAN[1,†], AND F. PUCCI[1,2,†]

## ABSTRACT

**Motivation:** The solubility of a protein is often decisive for its proper functioning. Lack of solubility is a major bottleneck in high-throughput structural genomic studies and in high-concentration protein production, and the formation of protein aggregates causes a wide variety of diseases. Since solubility measurements are time-consuming and expensive, there is a strong need for solubility prediction tools.
**Results:** We have recently introduced solubility-dependent distance potentials that are able to unravel the role of residue-residue interactions in promoting or decreasing protein solubility.Here, we extended their construction by defining solubility-dependent potentials based on backbone torsion angles and solvent accessibility, and integrated them, together with other structure- and sequence-based features, into a random forest model trained on a set of *E. coli* proteins with experimental structures andsolubility values. We thus obtained the SOLart protein solubility predictor, whose most informative features turned out to be folding free energy differences computed from our solubility-dependent statistical potentials. SOLart performances are very good, with a Pearson correlation coefficient between experimental and predicted solubility values of 0.7 both in the training dataset and on an independent set of *S. Cerevisiae* proteins. On test sets of modeled structures, only a limited drop in performance is observed. SOLart can thus be used with both high-resolution and low-resolution structures, and clearly outperforms state-of-art solubility predictors. It is available through a user-friendly webserver, which is easy to use by non-expert scientists.
**Availability:** The SOLart webserver is freely available at babylone.ulb.ac.be/SOLART/

## 1. Introduction

Solubility and aggregation are crucial properties of proteins, which can either ensure or prevent their correct functioning (Trevino *et al.*, 2008). Obtaining a thorough understanding of these matters is becoming increasingly important, since protein solubilization is required for improving a wide range of biotechnological and biopharmaceutical processes, especially when high protein concentrations are demanded. Just to mention some of them, protein solubility is frequently a serious bottleneck for the successful development of antibody therapeutics, which often suffer from aggregation at the conditions in which they are stored (Perchiacca and Tessier, 2012; Roberts, 2014), as well as in genome-wide structural analyses where about 80% of the total number of non-membrane proteins have been estimated to have insolubility-related problems (Golovanov *et al.*, 2004).

In the context of recombinant protein production, the formation of insoluble inclusion bodies, which are thought to contain clusters of different conformational states corresponding to folded, misfolded and partially folded structures, frequently makes the procedure to get bioactive proteins very challenging, involving first the solubilization of the inclusion bodies followed by the native refolding of the proteins (Martínez-Alonso *et al.*, 2009; Singh *et al.*, 2015; Baneyx and Mujacic, 2004; Singh and Panda, 2005; Vallejo and Rinas, 2004).

Scarse solubility properties are also directly related to pathological conditions such as the neurodegenerative Alzheimer and Parkinson diseases, whose hallmark is the progressive accumulation of insoluble deposits, *i.e* $\beta$-amyloid and $\alpha$-synuclein aggregates, respectively, that become toxic and interfere with the normal cell functioning (Chiti and Dobson, 2006; Bucciantini *et al.*, 2002; Irvine *et al.*, 2008; Ross and Poirier, 2004).

Reaching a full understanding of protein solubility mechanisms is particularly challenging, since solubility is a complex physicochemical property determined not only by various intrinsic factors such as residue-residue interactions, protein flexibility, amino acid composition and hydrophobicity, but also by various extrinsic variables such as the pH, the environmental temperature, the ionic strength and the protein concentration.

During the past decades, many efforts have been devoted to investigate the mechanisms and the factors that influence protein solubility (Trainor *et al.*, 2017). It has been reported that smaller proteins tend to have a higher solubility when overexpressed in *E. Coli* than longer proteins (Wilkinson and Harrison, 1991). The amino acid composition also influences protein solubility. For example, Asp, Glu and Ser contribute more favorably to solubility than other hydrophilic amino acids (Niwa *et al.*, 2009; Chan *et al.*, 2013); the values of the Lys/Arg and Glu/Asp ratios correlate with solubility (Warwicker *et al.*, 2013; Chan *et al.*, 2013); and aromatic-poor proteins tend to be more soluble than those enriched in aromatics (Niwa *et al.*, 2009; Hebditch *et al.*, 2017).

Furthermore, protein-protein and protein-solvent interactions have been shown to play key roles in the solubility properties. In particular, solvent exposed residues have some characteristics that are well correlated with solubility: insoluble proteins tend to have larger surface patches carrying a net positive charge than soluble proteins (Chan *et al.*, 2013), which are characterized instead by a more negatively charged surface (Kramer *et al.*, 2012).

In a recent work (Hou *et al.*, 2018), we showed that among all residue-residue interactions, the Lys-containing salt bridges and the aliphatic interactions contribute more strongly than others to promote solubility, whereas interactions involving delocalized $\pi$-electrons favor aggregation (*e.g.* aromatic, His-$\pi$, cation-$\pi$, amino-$\pi$ and anion-$\pi$ interactions). These different findings demonstrate the important potentiality of structural information in the understanding of the biophysical mechanisms underlying solubility data.

Several computational methods, mostly based on machine learning techniques, have recently been developed to predict protein solubility (Smialowski *et al.*, 2006, 2012; Idicula-Thomas *et al.*, 2005; Magnan *et al.*, 2009; Agostini *et al.*, 2014; Hebditch *et al.*, 2017; Khurana *et al.*, 2018; Hirose and Noguchi, 2013; Sormanni *et al.*, 2015). The large majority of the features that they use are extracted from the amino acid sequences, such as the sequence length, the amino acid composition, the absolute charge, the isoelectric point, the aliphatic index and the average hydropathy. Some other features are associated to structural properties, such as $\beta$-stand propensities or fractions of exposed and buried residues. However, these features are not assigned from the structure, but rather predicted from the sequence.

Structure-based techniques to predict solubility make use of extensive molecular dynamics simulations to evaluate the free energy difference between solution and aggregation phases (Tjong and Zhou, 2008). However these methods are computationally expensive and cannot be certainly applied to large-scale investigations of the protein solubility.

In summary, the current prediction methods tend to overlook structural data and require only the sequence as input. Clearly, considering features derived from experimental 3-dimensional (3D) structures adds valuable information, which should in principle boost the methods' performances. However, requiring 3D structures decreases the applicability of the predictor, as they are not always available. But this drawback is loosing importance, since homology modeling tools provide always better structural models that can safely be used by some predictors. Hence, progress is definitely expected in the solubility prediction field from the utilization of 3D structures.

In this paper, we fully exploited protein structure data through the use of statistical potentials, which have largely proven to be successful in many studies ranging from structure prediction to mutant analyses (see *e.g.*Kocher *et al.* (1994); Folch *et al.* (2010); Dehouck *et al.* (2009); Pucci *et al.* (2016)). More precisely, we used our recently developed solubility-dependent statistical potentials (Hou *et al.*, 2018) to discriminate between residue pair interactions that favor or disfavor protein solubility. In addition to these energetic features, we considered a series of other structure-based features and of commonly used sequence features. These were integrated into a predictor with the help of a random forest regression algorithm, so as to predict protein solubility with improved accuracy. Our predictor, called SOLart, is made freely available online at http://babylone.ulb.ac.be/SOLART/.

## 2. Materials and Methods

2.1. **Protein solubility definition.** We used as a definition of solubility $\mathcal{S}$ (in %) the ratio of the supernatant fraction obtained after centrifugation of the translation mixture over the total concentration of the overexpressed protein (Niwa *et al.*, 2009). It ranges from 0% to 130%. It is generally different from the physical solubility $\mathcal{S}_0$, measured in g/l and defined as the concentration of protein in a saturated solution that is in equilibrium with a solid phase.

$\mathcal{S}_0$ is difficult to measure and strongly depends on the type of precipitant used to perform the experiment and on the environmental variables such as the temperature. This makes the construction of a large dataset of $\mathcal{S}_0$ values for training and testing bioinformatic models almost impossible. We thus chose to use the $\mathcal{S}$ definition that can be measured and studied in large-scale investigations (Niwa *et al.*, 2009; Uemura *et al.*, 2018).

2.2. **Protein datasets.** To train and test SOLart, we considered two datasets of proteins that were expressed with the cell-free expression system called PURE (Shimizu *et al.*, 2005) and whose solubilities $\mathcal{S}$ were measured. These are Esol$_{Ecoli}$ and Esol$_{Scerevisiae}$, which contain the solubilities of about 70% of the entire *E. coli* K-12 strain proteome (Niwa *et al.*, 2009), and of around 500 cytosolic proteins from *S. cerevisiae* (Uemura *et al.*, 2018), respectively.

We used the functional and structural annotation server EcoGene (Zhou and Rudd, 2013) and the UNIPROT server (Apweiler *et al.*, 2004) to map the gene accession ids of every entry in these datasets onto the corresponding structures from the Protein Data Bank (PDB) (Berman *et al.*, 2000). Only X-ray structures with maximum 2.5 Å resolution, which have a sequence identity of 100% and at least 90% coverage with the associated Esol sequences, were selected.

We also considered the remaining proteins from the Esol datasets, which have no experimental structure. We collected structural models for these entries from the SWISS-MODEL repository (Bienert *et al.*, 2016). Only the models constructed using a template with a good resolution X-ray structure ($\leq$ 2.5 Å) and at least 30% sequence identity and 50% coverage with the query sequence were kept.

We first focused on the set of X-ray structures from *E. coli.* We used the protein-culling server PISCES (Wang and Dunbrack Jr, 2003) to select proteins with pairwise sequence identity of 25% at most. This dataset, called $\mathcal{D}_{Ecoli}$, contains 406 well resolved protein structures with experimental solubility values and low pairwise sequence identity. It is used as the SOLart training set.

Out of the modeled structures from *E. coli*, we dropped those that have a sequence identity of more than 40% with a protein from $\mathcal{D}_{Ecoli}$, and filtered out sequences with more than 40% pairwise identity. We obtained in this way the $\mathcal{M}_{Ecoli}$ dataset containing 679 protein models, which were used as a first test set.

We used the same procedure on the datasets of X-ray and modeled structures from *S. cerevisiae* proteins: we removed the entries that have more than 40% sequence identity with the training set $\mathcal{D}_{Ecoli}$, and finally filtered out proteins with more than 40% pairwise sequence identity. In this way, we obtained a third test set $\mathcal{D}_{Scerevisiae}$ composed of 70 X-ray structures and a fourth test set $\mathcal{M}_{Scerevisiae}$ with 64 structures obtained via homology modeling.

The proteins that are contained in the four datasets are listed in Tables S1 of Supplementary Material, with some additional information.

Our datasets could be suspected to be biased and to contain only some types of conformations. To check that they do not suffer from this problem, we mapped all structures from our datasets onto the corresponding CATH categories (Dawson *et al.*, 2016). As shown in Table 1, the 406 X-ray structures of $\mathcal{D}_{Ecoli}$ belong to 344 homologous superfamilies, which cover all four classes, 59% of the architectures and 15% of the folds. The three test datasets $\mathcal{D}_{Scerevisiae}$, $\mathcal{M}_{Ecoli}$ and $\mathcal{M}_{Scerevisiae}$ do not cluster into the same superfamilies. This indicates that the protein sets are unbiased, and not specifically enriched in certain superfamilies but tend to cover the full fold universe.

| | $\mathcal{D}_{Ecoli}$ (406) | | $\mathcal{M}_{Ecoli}$ (679) | | $\mathcal{D}_{Scerevisiae}$ (70) | | $\mathcal{M}_{Scerevisiae}$ (64) | |
|---|---|---|---|---|---|---|---|---|
| | n | % | n | % | n | % | n | % |
| C | 4 | 100 | 4 | 100 | 3 | 75 | 3 | 75 |
| A | 24 | 59 | 26 | 63 | 15 | 37 | 15 | 37 |
| T | 214 | 15 | 244 | 18 | 55 | 4 | 54 | 4 |
| H | 344 | 6 | 458 | 7 | 70 | 1 | 64 | 1 |

TABLE 1. **Mapping of the proteins of our datasets onto CATH categories (Dawson *et al.*, 2016).** The numbers in parentheses correspond to the amount of proteins in the corresponding dataset. 'C' stands for Class, 'A' for Architecture, 'T' for Topology or fold, and 'H' for Homologous superfamily. 'n' is the number of C, A, T or H categories that have at least member in the dataset and '%' the fraction of these categories represented in the dataset.

## 3. Results

3.1. **Features.** We used a series of features to set up the SOLart solubility predictor, which are described below.

- Statistical potentials

We applied and extended the solubility-dependent statistical potentials recently introduced in (Hou *et al.*, 2018), which have proven to yield an objective and informative description of the interactions that modulate protein solubility properties. The idea was to divide the dataset $\mathcal{D}_{Ecoli}$ into two subsets of equal size, called $\mathcal{D}^{\text{insol}}_{Ecoli}$ and $\mathcal{D}^{\text{sol}}_{Ecoli}$, which contain aggregation-prone and soluble proteins, respectively, and to derive distance potentials from each of the the two subsets (see (Hou *et al.*, 2018) for details). In this way, we defined two distinct potentials referred to as "insoluble" and "soluble".

Here we generalized this construction to other types of potentials involving various sequence and structure elements. In particular, for potentials based on one sequence element $s$ and one structure element $c$, we have:

$$\begin{aligned} \Delta W^{\text{insol}}(c,s) &= -k_B T \ln \frac{P(c,s,\mathcal{D}^{\text{insol}}_{Ecoli})}{P(s,\mathcal{D}^{\text{insol}}_{Ecoli})P(c,\mathcal{D}_{Ecoli})} \\ \Delta W^{\text{sol}}(c,s) &= -k_B T \ln \frac{P(c,s,\mathcal{D}^{\text{sol}}_{Ecoli})}{P(s,\mathcal{D}^{\text{sol}}_{Ecoli})P(c,\mathcal{D}_{Ecoli})} \end{aligned} \tag{1}$$

where $k_B$ is the Boltzmann constant and $T$ the absolute temperature. The sequence descriptor $s$ is an amino acid type, and the structure descriptor $c$ is either an inter-residue distance $d$ computed between the average geometric centers of the heavy side chain atoms, a backbone torsion angle domain $t$, or a solvent accessibility $a$ computed as the ratio between the solvent accessibility of a residue in a given structure and in an extended Gly-X-Gly tripeptide conformation (see *e.g.* (Kocher *et al.*, 1994; Rooman *et al.*, 1991; Pucci *et al.*, 2014)). $P(s, c, \mathcal{D})$ is the probability of joint occurrence of the sequence and structure elements $s$ and $c$ in the dataset $\mathcal{D}$, and similarly for the probability functions $P(s, \mathcal{D})$ and $P(c, \mathcal{D})$. These probabilities were estimated in terms of the number of occurrences of the sequence-structure associations in $\mathcal{D}$.

We constructed eleven solubility-dependent statistical potentials from different combinations of $s$ and $c$ elements, listed in Table 2. We named the potentials according to the type and number of sequence and structure descriptors. For example, "sa" represents the potential in which one amino acid type and one solvent accessibility are specified, whereas "sds" describes the potential in which two amino acid types and their interresidue distance are given.

We used these different potentials to compute folding free energy contributions of target proteins. As an example, the folding free energies computed from the soluble and insoluble ssd distance potentials were defined as:

$$\Delta G^{\alpha}_{\text{sds}} = \sum_{i=1}^{N} \sum_{j=i+2}^{N} \Delta W(s_i, s'_j, d_{ij}, \mathcal{D}^{\alpha}_{Ecoli}) \tag{2}$$

where $\alpha$ is equal to "sol" or "insol", $s_i$ and $s'_j$ are two residue types at positions $i$ and $j$ along the sequence, $d_{ij}$ is their spatial distance, and $N$ is the number of amino acids of the target protein. We then computed the folding free energy difference:

$$\Delta\Delta G_{\text{sds}} = \Delta G^{\text{sol}}_{\text{sds}} - \Delta G^{\text{insol}}_{\text{sds}} \tag{3}$$

Using analogous relations, we computed the folding free energy $\Delta G^{\alpha}$ and the folding free energy difference $\Delta\Delta G$ for each potential listed in Table 2.

- Protein size and accessible surface area

We considered three global characteristics of the proteins, which are the protein length ($\Lambda$), its solvent accessible surface area (SAcc) estimated with an in-house program (Dalkas *et al.*, 2014), and its solvent accessible surface area divided by the protein length (SAcc/$\Lambda$); in the latter case we used the length of the sequence whose structure has been determined. Note that the former feature is sequence-based, and that the latter two require the knowledge of the 3D structure.

- Secondary structure content

Another series of structure-based features were added, which are the fraction of protein residues that are in $\alpha$-helical, $\beta$-strand or coil (called here $\gamma$) conformation. We distinguished between the $\alpha$, $\beta$ and $\gamma$ residues that are buried in the protein core (solvent accessibility $\leq 20\%$), moderately buried (between 20 and 50%), and solvent exposed ($\geq 50\%$). Our in-house program (Dalkas *et al.*, 2014) was used to assign the secondary structure and solvent accessibility.

- Amino acid composition

We integrated 20 purely sequence-based features, corresponding to the fraction of each of the 20 amino acid present in a protein. We also considered the fraction of amino acid groups, *i.e.* positively charged residues (K+R), negatively charged residues (D+E), charged residues (K+R+D+E) aromatic residues (F+W+Y), as well as the difference between the fractions of K and R (K-R), D and E (D-E), and K+R and D+E (K+R-D-E). We combined these features with the solvent accessibility and defined three categories per amino acid or amino acid group, according to whether the residue is exposed, moderately buried or buried. This yielded 81 additional structure-based features.

3.2. **Feature selection.** The next step consisted in selecting, out of the above-defined 28 purely sequence-based features and 103 structure-based features, the subset of features that are the most informative for protein solubility. We used for that purpose the $\mathcal{D}_{Ecoli}$ training set, which contains 406 non-redundant high-resolution X-ray structures of *E. coli* proteins with low pairwise sequence identity and experimentally measured solubility (see Methods). The feature selection was performed using the Boruta algorithm (Kursa *et al.*, 2010) implemented in the Caret package of R (Kuhn *et al.*, 2008), a wrapper built around the random forest classification algorithm (Liaw *et al.*, 2002), which compares the importance of the real features with those of random (shadow) features using statistical testing. The results are obtained as an average over several runs (here 1,000) of random forest.

We filtered out the features whose average importance measured by the Boruta algorithm is lower than 1. This led us to keep a total of 52 features, which are shown in in Fig. 1 and Supplementary Information Fig. S1. Among these, 37 require the knowledge of the structure.

Strikingly, the four top-ranked features are folding free energy differences $\Delta\Delta G$ computed from our solubility-dependent potentials: the backbone torsion angle potential sst, the solvent accessibility potential ssa and the two distance potentials sd and sds (see Table 2). The next most important feature is the protein length $\Lambda$, followed by the solvent accessibility and fractions of some amino acid types. The features based on the secondary structure do not appear among the 30 top features, but some appear in the list of 52 selected features.

3.3. **Setting up SOLart.** The 52 selected features were combined to set up the SOLart predictor of the solubility of target proteins on the basis of their 3D structures. We used for that purpose $\mathcal{D}_{Ecoli}$ as training set, and the random forest regression algorithm implemented in the R package (Liaw *et al.*, 2002) to construct the model. This algorithm is a tree-based system composed of multiple regression trees; the number of trees is here set to 500. The training process starts with a randomly selected subset of the original dataset from which a regression tree is constructed by the iterative partitioning of the data space into smaller subsets. At each node of the tree, randomly sampled features are used; the number of features depends on a global parameter ”mtry” taken here between 1 and 3. The optimal mtry value is obtained through a grid search procedure as the one that yields the highest correlation coefficient in the training dataset. The regression for a target protein is obtained by averaging the predictions over all trees.

3.4. **Performance of SOLart.** SOLart's performances were evaluated by three replicates of a 10-fold cross validation procedure on the $\mathcal{D}_{Ecoli}$ training set. The replicates were performed with different random divisions into folds, and the performances were computed as averages over the replicates. Our computational model reaches a good linear correlation coefficient $r = 0.67$ between the SOLart solubility predictions and the experimental values, and a root mean square error RMSE= 25% (Table 3).

We also tested SOLart on an independent test set that contains *S. cerevisiae* proteins with a well resolved X-ray structure, grouped in the $\mathcal{D}_{Scerevisiae}$ set (see Methods). The performance of SOLart on this set is evaluated by a linear correlation coefficient $r = 0.70$ and an RMSE = 23%. When 10% outliers are removed, the score increases up to $r = 0.78$ and RMSE = 19% (Table 3). The scores on this independent set are thus even better than those obtained in cross validation on the training set $\mathcal{D}_{Ecoli}$.

To further analyze this result, we estimated the importance of each feature in the SOLart prediction using the varImp permutation scheme-based function (Kuhn *et al.*, 2008). It proceeds by randomly permuting each feature in turn in order to break its association with the response, and then using it together with the remaining unpermuted features for prediction. The decrease of the prediction accuracy is a measure of the importance of the permuted feature. This measure estimates the weight of each individual feature in the predictor, whereas the feature selection algorithm applied in section 3.2 measures the feature relevance independently of the prediction model. They thus yield slightly different rankings.

The 20 most important features of our prediction model are shown in Fig. 3. Interestingly, almost all the features that correspond to folding free energy differences ($\Delta\Delta G$) are in this list (9 out of 11), and the 6 top features are the $\Delta\Delta G$s computed from the potentials ssa, sst, sd, sds, saa, and sa (Table 2). The two best ones, almost *ex æquo*, are $\Delta\Delta G_{\text{ssa}}$ and $\Delta\Delta G_{\text{sst}}$, which also ranked first in the feature selection (Fig. 1). They are computed from the propensities of amino acid pairs to be associated with a certain solvent accessibility range $a$ or a certain backbone torsion angle domain $t$ of a residue. These propensities differ between soluble and aggregation-prone proteins, and it is this difference which is measured through the $\Delta\Delta G$ features. The next best ranked features are $\Delta\Delta G_{\text{sd}}$ and $\Delta\Delta G_{\text{sds}}$, computed from the propensities of residue pairs to be separated by a certain spatial distance, followed by two other accessibility potentials $\Delta\Delta G_{\text{saa}}$ and $\Delta\Delta G_{\text{sa}}$.

These folding free energy features require the protein structure as input. In fact, more than half of the top 20 features are structure-based. This confirms the relevance of the structural information in the determination of the protein solubility properties. The first sequence-based feature ranks seventh. It is the sequence length $\Lambda$: in general, the smaller the sequence, the most soluble the protein (Kramer *et al.*, 2012). The two related features, *i.e.* the solvent accessible surface area SAcc divided or not by the length, are also among the top 20 features.

The remaining top 20 features are sequence-based: the difference between Lys and Arg composition (K-R) which is positively correlated with solubility (Warwicker *et al.*, 2013; Hou *et al.*, 2018), the percentage of aromatic residues (F+Y+W) which favor aggregation (Niwa *et al.*, 2009; Hou *et al.*, 2018), and the total fraction of negatively charged residue (D+E) that have also been shown to promote solubility (Niwa *et al.*, 2009; Hou *et al.*, 2018). The next features are the composition in R and Q, which disfavors solubility, the composition in E and K, which instead promotes solubility, and the difference between the fraction of positively and negatively charged residues (K+R-D-E), which augments insolubilty.

Note that all these sequence-based features are also employed by the solubility predictors available in the literature. However, in addition to these commonly used features, we exploit a series of structure-based features among which the most important ones are obtained from the newly developed solubility-dependent statistical potentials. These capture the solubility properties in a more accurate way and represent the key instrument of our approach.

3.5. **Performance on modeled protein structures.** SOLart has been shown to be accurate when the 3D structure of the target protein is known. To enlarge its applicability, we tested it on low-resolution structures obtained via homology modeling. We first applied it to the $\mathcal{M}_{Ecoli}$ dataset containing 679 proteins from *E. coli* (see Methods). We obtained a correlation of $r = 0.51$ and a RMSE of 28%, which is relatively good but lower than the performance on $\mathcal{D}_{Ecoli}$ (Table 3). This drop is expected since we have to take into account the possible inaccuracies in the modeled structures that have to be added to the error of our computational method. After removing 10% outliers, the performance increases to $r = 0.66$ and RMSE= 23%, and reaches thus the same performance as on good-resolution structures.

As a last test set, we used $\mathcal{M}_{Scerevisiae}$ that contains *S. cerevisiae* proteins with modeled structures. The performance of SOLart on this set is given by $r = 0.65$ and RMSE = 24%, and increases up to $r = 0.71$ and RMSE = 20% without 10% outliers. The scores are thus much higher on this test set than on the *E. coli* test set, which suggests that some structural protein models or experimental solubility values might be less accurate on the the *E. coli* set than on the *S. cerevisiae* set.

Note that these tests are quite strict, since there is a low sequence similarity (¡40%) between these test sets and the training set. We thus conclude that SOLart can reliably be used to predict solubility not only for high-resolution experimental structures but also for modeled or other low-resolution structures.

3.6. **Comparison with other solubility prediction methods.** The performance of SOLart was compared with that of other solubility prediction methods on the combination of $\mathcal{D}_{Scerevisiae}$ and $\mathcal{M}_{Scerevisiae}$ sets, that group X-ray and modeled structures from *S. cerevisiae* proteins, as these are independent test sets that are not included in the training sets of any of the predictors. More precisely, we tested the methods Protein-SOL (Hebditch *et al.*, 2017), ccSOL (Agostini *et al.*, 2014), CamSol (Sormanni *et al.*, 2015), PROSO (Smialowski *et al.*, 2006), PROSO II (Smialowski *et al.*, 2012) and SOLpro (Magnan *et al.*, 2009), by submitting to their respective webservers all the proteins from our test datasets. Note that all these methods are sequence-based.

The linear correlation coefficient $r$ between the solubility predictions and the experimental values for all these predictors are given in Table 4. Our method clearly outperforms the competitors ($r = 0.68$ against $r = 0.56$ for the second best method). This demonstrates the importance of using structural information.

3.7. **Webserver.** We provided a freely available webserver interface for our prediction method, which targets non-expert users (http://babylone.ulb.ac.be/ SOLART/index.php) (Fig. 4). The input consists of the 3D structure of the target protein in PDB format. It can be uploaded directly by the user or imported from the PDB (Berman *et al.*, 2000) by typing its 4-letter code. The webserver then provides a brief summary of some of the protein's characteristics and allows the user to choose one of the protein chains. The computation starts after the query submission. All the structure-based free energy, secondary structure and solvent accessibility features are first computed and then integrated with the other, sequence-based, features.

In the output page, reached by following the link provided, the value of the predicted scaled solubility $\mathcal{S}$ is given. If the score is close to zero, the target protein is predicted as aggregation-prone and, when it is close to 130, as soluble. Moreover, to have an indication of the contribution of each single feature to the solubility prediction of the target protein, we also show a figure (Fig. 4) with the solubility predicted from each feature taken individually and with all SOLart features together. The prediction with each single feature is computed from a random forest model trained on the experimental solubility values of the $\mathcal{D}_{Ecoli}$ set. This figure can be used as a source of inspiration to suggest the characteristics to modify in view of modulating solubility.

Due to its simplicity of use, we expect that this webserver will be of interest for researchers in academia and industry who are interested in modulating protein solubility without needing any prior bioinformatic knowledge.

## 4. Discussion

We introduced SOLart as on of the first structure-based solubility prediction method, which is able to predict quickly and accurately the protein solubility of a protein from its experimental or modeled 3D structure.

SOLart employs a series of features, among which the sequence-based features that are commonly used for solubility prediction and some classical structure-based features such as secondary structure composition and solvent accessibility. In addition, it takes advantage of the potentiality of solubility-dependent statistical potentials to discriminate the residue interactions that favor aggregation or solubility. Besides the distance potentials that have previously been analyzed (Hou *et al.*, 2018), ten new solubility-dependent potentials were introduced here, which describe the local propensities of residues to adopt certain backbone torsion angle domains or to have certain solvent accessibility values in soluble or aggregation-prone proteins. Note that the feature importance analyses show that the torsion, solvent accessibility and distance potentials are the most important features in the random forest regression prediction. The folding free energy differences computed with these potentials are better correlated with solubility than other protein properties analyzed in the literature such as protein length, isoelectric point and aliphatic index.

The performances of SOLart are high and robust: the linear correlation on both the training dataset (in cross validation) and on three independent test sets almost reaches 0.7 on good-resolution X-ray structures and slightly lower on modeled structures. It is important to underline that SOLart can be used with modeled structures, as it largely expands the domain of applicability of our tool. Furthermore, it performs similarly in the training and testing datasets, which indicates its robustness and absence of bias towards the training set. Finally, SOLart outperforms the state-of-the-art solubility predictors on an independent dataset containing *S. Cerevisiae* proteins, with an increase of more thant 20% in the correlation coefficient between the predicted and the experimental values of the solubility. This provides a strong demonstration of SOLart's accuracy and usefulness.

Another advantage of SOLart is its fastness: it is able to predict the solubility of a medium-size protein in less than one minute. This quality make this tool a perfect instrument to investigate protein solubilty properties on a large scale.

Even though SOLart performances are good, there is still a lot of work needed to unravel the various effects and to understand the biophysical mechanisms underlying solubility and aggregation. One direction is to design better energy functions that describe more efficiently these phenomena by enlarging the protein datasets with experimental solubility values or modifying their original formulation. For example, the definition of the reference state that is adequate for solubility properties is still an open problem. It has been argued that interactions between unfolded conformations could lead to insoluble aggregates and, indeed, inclusion bodies forming in heterologous expression in *E. coli* have been shown to involve folded, unfolded, misfolded and partially folded proteins (Martínez-Alonso *et al.*, 2009; Singh *et al.*, 2015; Baneyx and Mujacic, 2004; Singh and Panda, 2005; Vallejo and Rinas, 2004), which makes it challenging to disentangle the characteristics contributing to its formation.

Note also that the definition of the solubility ($\mathcal{S}$) used in this paper differs from the physical definition of solubility ($\mathcal{S}_0$), measured in g/l, defined as the concentration of a protein in a saturated solution that is in equilibrium with a solid phase. To get insights into the relation between these two solubility definitions, they should systematically be compared. This is currently impossible as no large datasets of $\mathcal{S}_0$ values are available due to the difficulties in its experimental measurements.

A final perspective concerns industrial biotechnological applications, in which water is replaced by other polar solvents or even by non-polar solvents. Understanding how the protein solubility changes according to the type of solvent and being able to accurately predict this change is a major target for computational tools. On the same footing, it would also be important to understand and predict the influence of buffer salts and ionic strength on the solubility properties of proteins.

In summary, SOLart is a new and efficient method to predict protein solubility. Thanks to its user-friendly interface, both expert and non-expert users can use its predictions to analyze and improve the solubility properties of targeted proteins involved in biotechnological processes, where solubility is frequently a major bottleneck.

## Funding

This work is supported by the FNRS Fund for Scientific Research through a PDR grant; Q.H. and F.P. are Postdoctoral Researchers and M.R. is Research Director at the FNRS.

F.P. has been partially supported by the John von Neumann Institute for Computing (NIC).

## References

Agostini,F., Cirillo,D., Livi,C.M., Delli Ponti,R. and Tartaglia,G.G. (2014) cc sol omics: a webserver for solubility prediction of endogenous and heterologous expression in escherichia coli. *Bioinformatics,* **30** (20), 2975–2977.

Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) Uniprot: the universal protein knowledgebase. *Nucleic acids research,* **32** (suppl_1), D115–D119.

Baneyx,F. and Mujacic,M. (2004) Recombinant protein folding and misfolding in escherichia coli. *Nature biotechnology,* **22** (11), 1399.

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res,* **28** (1), 235–242.

Bienert,S., Waterhouse,A., de Beer,T.A., Tauriello,G., Studer,G., Bordoli,L. and Schwede,T. (2016) The swiss-model repository—new features and functionality. *Nucleic acids research,* **45** (D1), D313–D319.

Bucciantini,M., Giannoni,E., Chiti,F., Baroni,F., Formigli,L., Zurdo,J., Taddei,N., Ramponi,G., Dobson,C.M. and Stefani,M. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *nature,* **416** (6880), 507.

Chan,P., Curtis,R.A. and Warwicker,J. (2013) Soluble expression of proteins correlates with a lack of positively-charged surface. *Scientific Reports,* **3**, 3333.

Chiti,F. and Dobson,C.M. (2006) Protein misfolding, functional amyloid, and human disease. *Annu. Rev. Biochem.,* **75**, 333–366.

Dalkas,G.A., Teheux,F., Kwasigroch,J.M. and Rooman,M. (2014) Cation–$\pi$, amino–$\pi$, $\pi$–$\pi$, and h-bond interactions stabilize antigen–antibody interfaces. *Proteins: Structure, Function, and Bioinformatics,* **82** (9), 1734–1746.

Dawson,N.L., Lewis,T.E., Das,S., Lees,J.G., Lee,D., Ashford,P., Orengo,C.A. and Sillitoe,I. (2016) Cath: an expanded resource to predict protein function through structure and sequence. *Nucleic acids research,* **45** (D1), D289–D295.

Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: popmusic-2.0. *Bioinformatics,* **25** (19), 2537–2543.

Folch,B., Dehouck,Y. and Rooman,M. (2010) Thermo-and mesostabilizing protein interactions identified by temperature-dependent statistical potentials. *Biophysical journal,* **98** (4), 667–677.

Golovanov,A.P., Hautbergue,G.M., Wilson,S.A. and Lian,L.Y. (2004) A simple method for improving protein solubility and long-term stability. *Journal of the American Chemical Society,* **126** (29), 8933–8939.

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** (19), 3098–3100.

Hirose,S. and Noguchi,T. (2013) Espresso: a system for estimating protein expression and solubility in protein expression systems. *Proteomics,* **13** (9), 1444–1456.

Hou,Q., Bourgeas,R., Pucci,F. and Rooman,M. (2018) Computational analysis of the amino acid interactions that promote or decrease protein solubility. *Scientific reports,* **8** (1), 14661.

Idicula-Thomas,S., Kulkarni,A.J., Kulkarni,B.D., Jayaraman,V.K. and Balaji,P.V. (2005) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in escherichia coli. *Bioinformatics,* **22** (3), 278–284.

Irvine,G.B., El-Agnaf,O.M., Shankar,G.M. and Walsh,D.M. (2008) Protein aggregation in the brain: the molecular basis for alzheimer's and parkinson's diseases. *Molecular medicine,* **14** (7-8), 451.

Khurana,S., Rawi,R., Kunji,K., Chuang,G.Y., Bensmail,H. and Mall,R. (2018) Deepsol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics,* **34** (15), 2605–2613.

Kocher,J.P.A., Rooman,M.J. and Wodak,S.J. (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *Journal of molecular biology,* **235** (5), 1598–1613.

Kramer,R.M., Shende,V.R., Motl,N., Pace,C.N. and Scholtz,J.M. (2012) Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophysical journal,* **102** (8), 1907–1915.

Kuhn,M. *et al.* (2008) Building predictive models in r using the caret package. *Journal of statistical software,* **28** (5), 1–26.

Kursa,M.B., Rudnicki,W.R. *et al.* (2010) Feature selection with the boruta package. *J Stat Softw,* **36** (11), 1–13.

Liaw,A., Wiener,M. *et al.* (2002) Classification and regression by randomforest. *R news,* **2** (3), 18–22.

Magnan,C.N., Randall,A. and Baldi,P. (2009) Solpro: accurate sequence-based prediction of protein solubility. *Bioinformatics,* **25** (17), 2200–2207.

Martínez-Alonso,M., González-Montalbán,N., García-Fruitós,E. and Villaverde,A. (2009) Learning about protein solubility from bacterial inclusion bodies. *Microbial cell factories,* **8** (1), 4.

Niwa,T., Ying,B.W., Saito,K., Jin,W., Takada,S., Ueda,T. and Taguchi,H. (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. *Proceedings of the National Academy of Sciences,* **106** (11), 4201–4206.

Perchiacca,J.M. and Tessier,P.M. (2012) Engineering aggregation-resistant antibodies. *Annual review of chemical and biomolecular engineering,* **3**, 263–286.

Pucci,F., Bourgeas,R. and Rooman,M. (2016) Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing hotmusic. *Scientific reports,* **6**, 23257.

Pucci,F., Dhanani,M., Dehouck,Y. and Rooman,M. (2014) Protein thermostability prediction within homologous families using temperature-dependent statistical potentials. *PLoS One,* **9** (3), e91659.

Roberts,C.J. (2014) Therapeutic protein aggregation: mechanisms, design, and control. *Trends in biotechnology,* **32** (7), 372–380.

Rooman,M.J., Kocher,J.P.A. and Wodak,S.J. (1991) Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *Journal of molecular biology,* **221** (3), 961–979.

Ross,C.A. and Poirier,M.A. (2004) Protein aggregation and neurodegenerative disease. *Nature medicine,* **10** (7), S10.

Shimizu,Y., Kanamori,T. and Ueda,T. (2005) Protein synthesis by pure translation systems. *Methods,* **36** (3), 299–304.

Singh,A., Upadhyay,V., Upadhyay,A.K., Singh,S.M. and Panda,A.K. (2015) Protein recovery from inclusion bodies of escherichia coli using mild solubilization process. *Microbial cell factories,* **14** (1), 41.

Singh,S.M. and Panda,A.K. (2005) Solubilization and refolding of bacterial inclusion body proteins. *Journal of bioscience and bioengineering,* **99** (4), 303–310.

Smialowski,P., Doose,G., Torkler,P., Kaufmann,S. and Frishman,D. (2012) Proso ii–a new method for protein solubility prediction. *The FEBS journal,* **279** (12), 2192–2200.

Smialowski,P., Martin-Galiano,A.J., Mikolajka,A., Girschick,T., Holak,T.A. and Frishman,D. (2006) Protein solubility: sequence based prediction and experimental verification. *Bioinformatics,* **23** (19), 2536–2542.

Sormanni,P., Aprile,F.A. and Vendruscolo,M. (2015) The camsol method of rational design of protein mutants with enhanced solubility. *Journal of molecular biology,* **427** (2), 478–490.

Tjong,H. and Zhou,H.X. (2008) Prediction of protein solubility from calculation of transfer free energy. *Biophysical journal,* **95** (6), 2601–2609.

Trainor,K., Broom,A. and Meiering,E.M. (2017) Exploring the relationships between protein sequence, structure and solubility. *Current opinion in structural biology,* **42**, 136–146.

Trevino,S.R., Scholtz,J.M. and Pace,C.N. (2008) Measuring and increasing protein solubility. *Journal of pharmaceutical sciences,* **97** (10), 4155–4166.

Uemura,E., Niwa,T., Minami,S., Takemoto,K., Fukuchi,S., Machida,K., Imataka,H., Ueda,T., Ota,M. and Taguchi,H. (2018) Large-scale aggregation analysis of eukaryotic proteins reveals an involvement of intrinsically disordered regions in protein folding. *Scientific reports,* **8** (1), 678.

Vallejo,L.F. and Rinas,U. (2004) Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microbial cell factories,* **3** (1), 11.

Wang,G. and Dunbrack Jr,R.L. (2003) Pisces: a protein sequence culling server. *Bioinformatics,* **19** (12), 1589–1591.

Warwicker,J., Charonis,S. and Curtis,R.A. (2013) Lysine and arginine content of proteins: computational analysis suggests a new tool for solubility design. *Molecular pharmaceutics,* **11** (1), 294–303.

Wilkinson,D.L. and Harrison,R.G. (1991) Predicting the solubility of recombinant proteins in escherichia coli. *Nature Biotechnology,* **9** (5), 443.

Zhou,J. and Rudd,K.E. (2013) EcoGene 3.0. *Nucleic Acids Research,* **41** (Database-Issue), 613–624.

[1]Computational Biology and Bioinformatics, CP 165/61, Université Libre de Bruxelles, 50 Roosevelt Ave, 1050 Brussels, Belgium;

[2]John von Neumann Institute for Computing, Jülich Supercomputer Centre, Forschungszentrum Jülich, 52428 Jülich, Germany

† Contributed equally to this work

*E-mail address*: mrooman@ulb.ac.be

| Features | Description | SOLart |
|---|---|---|
| **Statistical potentials** | | |
| sd: $\Delta\Delta G_{\mathrm{sd}}$ | 1 amino acid, 1 distance | ✓✓ |
| sds: $\Delta\Delta G_{\mathrm{sds}}$ | 2 amino acids, 1 distance | ✓✓ |
| sa: $\Delta\Delta G_{\mathrm{sa}}$ | 1 amino acid, 1 solvent accessibility | ✓✓ |
| saa: $\Delta\Delta G_{\mathrm{saa}}$ | 1 amino acid, 2 solvent accessibilities | ✓✓ |
| ssa: $\Delta\Delta G_{\mathrm{ssa}}$ | 2 amino acids, 1 solvent accessibility | ✓✓ |
| st: $\Delta\Delta G_{\mathrm{st}}$ | 1 amino acid, 1 torsion angle domain | ✓✓ |
| stt: $\Delta\Delta G_{\mathrm{stt}}$ | 1 amino acid, 2 torsion angle domains | ✓✓ |
| sst: $\Delta\Delta G_{\mathrm{sst}}$ | 2 amino acids, 1 torsion angle domain | ✓✓ |
| sad: $\Delta\Delta G_{\mathrm{sad}}$ | 1 amino acid, 1 distance and 1 solvent accessibility | ✓✓ |
| std: $\Delta\Delta G_{\mathrm{std}}$ | 1 amino acid, 1 distance and 1 torsion angle domain | ✓✓ |
| sta: $\Delta\Delta G_{\mathrm{sta}}$ | 1 amino acid, 1 distance and 1 solvent accessibility | ✓✓ |
| **Protein size and solvent accessible surface area** | | |
| $\Lambda$ | protein length | ✓✓ |
| SAcc | protein solvent accessibility | ✓✓ |
| SAcc/$\Lambda$ | protein solvent accessibility divided by length | ✓✓ |
| **Secondary structure content** | | |
| $\beta$ _b | fraction of buried $\beta$ residues | ✓✓ |
| $\beta$_m | fraction of moderately buried $\beta$ residues | ✓✓ |
| $\beta$_e | fraction of exposed $\beta$ residues | |
| $\alpha$_b | fraction of buried $\alpha$ residues | |
| $\alpha$_m | fraction of moderately buried $\alpha$ residues | ✓✓ |
| $\alpha$_e | fraction of exposed $\alpha$ residues | ✓✓ |
| $\gamma$_b | fraction of buried coil residues | |
| $\gamma$_m | fraction of moderately buried coil residues | |
| $\gamma$_e | fraction of exposed coil residues | |
| **Amino acid composition** | | |
| $\mathrm{C}_{i\ (i=1..20)}$ | fraction of each of the 20 amino acid types | ✓ |
| K+R | fraction of positively charged residues | |
| K-R | fraction of K minus fraction of R | ✓✓ |
| D+E | fraction of negatively charged residues | ✓✓ |
| D-E | fraction of D minus fraction of E | |
| K+R+D+E | fraction of charged residues | ✓✓ |
| K+R-D-E | fraction of positively minus negatively charged residues | ✓✓ |
| F+W+Y | fraction of aromatic residues | ✓✓ |
| _b, m, e | *idem* with distinction between buried, moderately buried and exposed residues | ✓ |

Table 2. **List of all the features tested for SOLart.** Those used in the final version are marked by a ✓✓; those for which a subset is used are marked by a ✓.

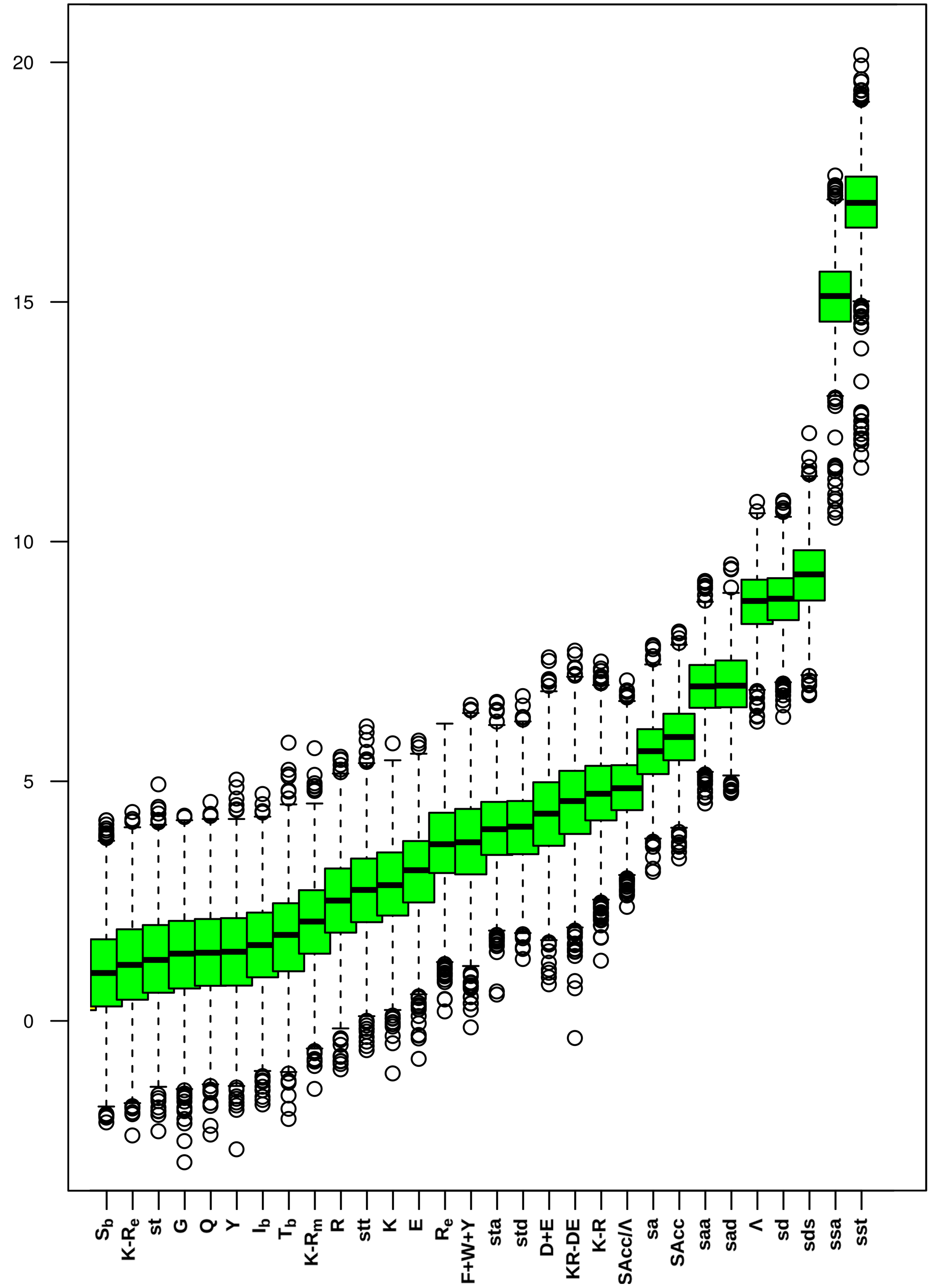


FIGURE 1. **The top 30 most important features identified by feature selection, from left to right.** The names in lower-case letters indicate folding free energy differences, *e*.g sst means $\Delta\Delta G_{\text{sst}}$.

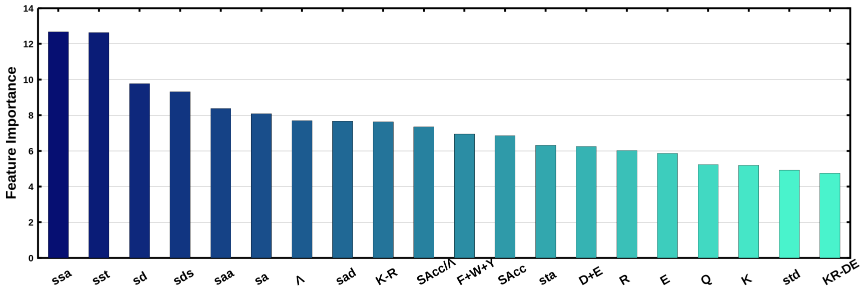


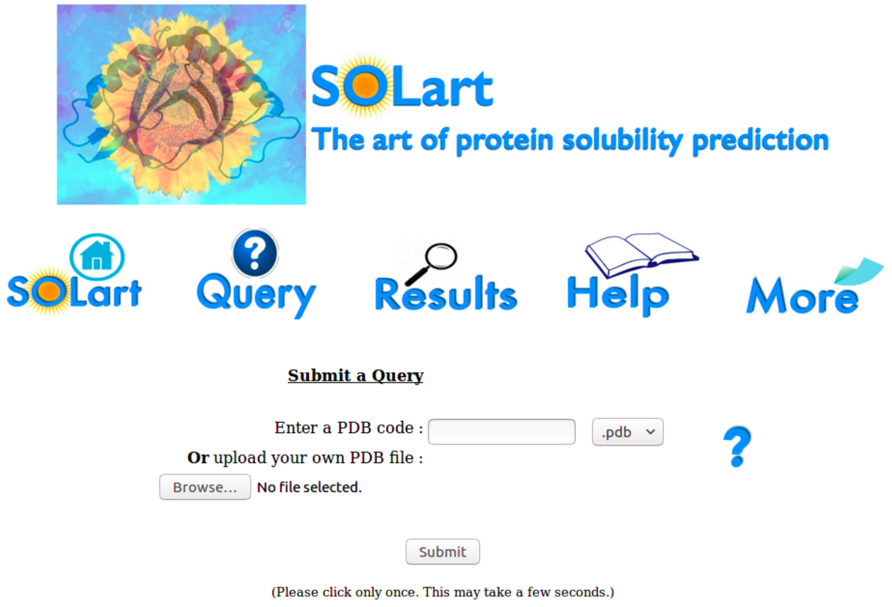
Figure 2. **The top 20 most important features of SOLart, from right to left.** The names in lower-case letters indicate folding free energy differences, *e*.g sst means $\Delta\Delta G_{\text{ssa}}$.

| | $\mathcal{D}_{Ecoli}$ | $\mathcal{M}_{\text{Ecoli}}$ | $\mathcal{D}_{Scerevisiae}$ | $\mathcal{M}_{Scerevisiae}$ |
|---|---|---|---|---|
| $r$ | 0.67 | 0.51 (0.66) | 0.70 (0.78) | 0.65 (0.71) |
| RMSE | 25% | 28% (23%) | 23% (19%) | 24% (20%) |

Table 3. **SOLart performances** in cross validation on the learning set $\mathcal{D}_{Ecoli}$, and on three independent test sets: $\mathcal{D}_{Scerevisiae}$ containing X-ray structures and $\mathcal{M}_{Ecoli}$ and $\mathcal{M}_{Scerevisiae}$ containing modeled structures. The values in parentheses correspond to the performance with 10% outliers removed.

| Predictor | $r$ |
|---|---|
| SOLart | 0.68 |
| Protein-Sol | 0.56 |
| ccSOL | 0.56 |
| CamSol | 0.39 |
| PROSO II | 0.13 |
| SOLpro | 0.18 |
| PROSO | 0.20 |

Table 4. **Comparison of the performance of different predictors on the combination of the $\mathcal{D}_{Scerevisiae}$ and $\mathcal{M}_{Scerevisiae}$ test sets,** on the basis of the Pearson correlation coefficient between predicted and experimental solubility values. SOLpro and PROSO first predict the proteins as soluble or insoluble and then give a probability score; we thus calculated the correlation by considering the solubility to be -1 for proteins predicted as insoluble and +1 for proteins predicted as soluble.



**Submit a Query**

Enter a PDB code :

**Or** upload your own PDB file :

No file selected.

(Please click only once. This may take a few seconds.)

FIGURE 3. **The webserver interface of SOLart.**

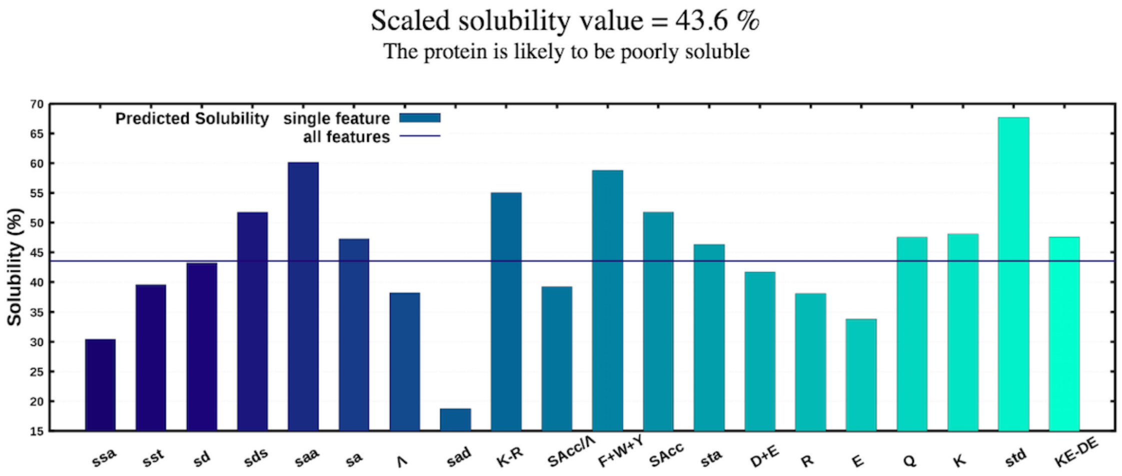


FIGURE 4. **Predicted solubility of an example protein (PDB code 2qia, Uniprot code P0A722) with all features used in SOLart (horizontal line) or with each single feature only (histogram bars).**