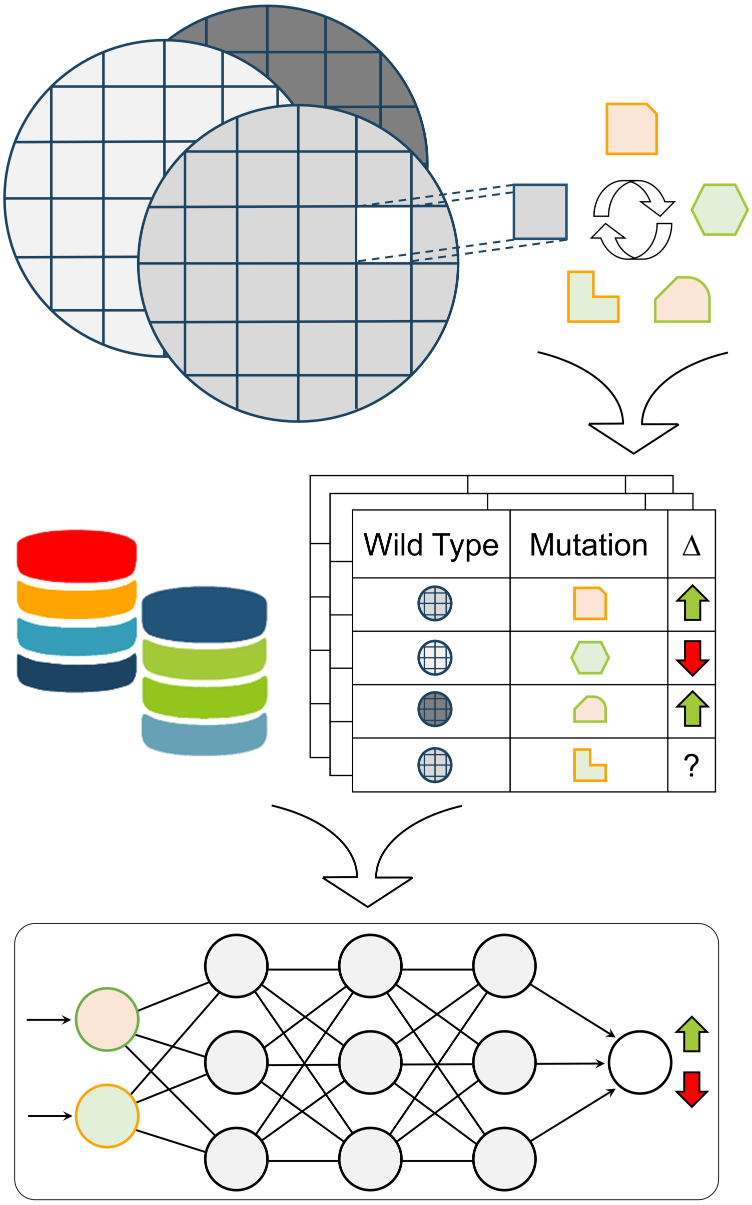
doi.org/10.1002/cctc.202000933

# Predicting protein stability and solubility changes upon mutations: data perspective

Stanislav Mazurenko*[a]



Understanding mutational effects on protein stability and solubility is of particular importance for creating industrially relevant biocatalysts, resolving mechanisms of many human diseases, and producing efficient biopharmaceuticals, to name a few. For *in silico* predictions, the complexity of the underlying processes and increasing computational capabilities favor the use of machine learning. However, this approach requires sufficient training data of reasonable quality for making precise predictions. This minireview aims to summarize and scrutinize available mutational datasets commonly used for training predictors. We analyze their structure and discuss the possible directions of improvement in terms of data size, quality, and availability. We also present perspectives on the development of mutational data for accelerating the design of efficient predictors, introducing two new manually curated databases FireProt$^{DB}$ and SoluProtMut$^{DB}$ for protein stability and solubility, respectively.

## 1. Introduction

Efficient design of stable and soluble protein variants is one of the principal goals of biocatalyst engineering. Understanding the mechanisms governing protein stability and solubility changes upon mutations is of paramount importance in several domains, including biotechnology, medicine, and biopharmaceutics.[1,2] Biocatalyst production suffers losses from time and resources wasted on poorly soluble and unstable protein mutants,[3] and in industrial applications, improved stability against harsh environments often becomes critical.[4] Improving properties of valuable but difficult-to-work-with proteins that have borderline stability or are poorly soluble outside living cells presents another major challenge in biotechnology.[5,6] Moreover, human neurodegenerative disorders, metabolic diseases, and cancer are often linked to mutations leading to protein misfolding, aggregation,[7–9] or decreased solubility,[10,11] and unstable or insoluble proteins may lead to precipitates triggering an unwanted immune response in patients.[12]

Given the astoundingly vast protein sequence space to explore in the pursuit of improved stability and solubility, computational tools are used increasingly to narrow down the search to ideally only a few promising mutations to be tested experimentally.[1] Many recent successes in designing improved variants relied on incorporating *in silico* prediction in the pipeline,[13,14] e.g. in the recent application of computer-assisted protein engineering strategy to modify fibroblast growth factor[15] to yield unprecedented stability and uncompromised biological function (Figure 1). Several new reviews provide excellent overviews of modern approaches and success stories in applying computational methods for engineering stable and soluble biocatalysts.[16–20]

Computational tools often rely on a set of rationally chosen rules applied at different steps of the protein engineering workflow. Many factors may potentially affect the outcome of introducing mutations: from physico-chemical properties of substitutions locally[21] to global changes in the protein backbone.[22] Composite combinations of those factors have also been reported to lead to better design capabilities, e.g. protein spectra derived by digital signal processing was shown to be useful in the design of stereoselectivity.[23,24] This complexity promotes the use of machine learning (ML) techniques, i.e. general-purpose algorithms for automatic rule generation based on patterns in available data.[25] Such algorithms have already substantially advanced our capabilities in image analysis, speech recognition, natural language processing, and other intrinsically complex tasks.[26,27] Therefore, their application to such sophisticated problems as predicting mutational effects on protein stability and solubility was only a matter of time.

Many promising ML-based predictors have been published for either task.[1,28–30] However, they all seem to be testing a similar limit to the prediction accuracy, e.g. the root mean square error of around 1 kcal/mol for stability predictions.[1] Moreover, independent experimental validation in subsequent studies often reveals modest performance.[28,31–33] Several explanations can be provided, one of which is the limited data size and quality available for training: data quality and abundance are critical for ML algorithms as they ultimately aim to identify and generalize patterns in the training data.

In this minireview, we focus on the databases and data sets habitually used to train such predictors. We briefly discuss their structure and associated challenges, from misleading notations and erroneous entries to the problem of aggregating results from different experimental setups. We conclude with the perspectives in improving those sets with the hope that it will further accelerate the usage of modern data analysis approaches to uncover the driving forces behind the effects in question.

We chose to consider both stability and solubility mutational data due to the similar structure of data as well as intertwined effects of the underlying mechanisms. On the one hand, unstable proteins tend to aggregate and are prone to faster degradation by proteolysis,[34] producing a negative signal in solubility assays. On the other hand, protein stabilization achieved by means of protein engineering was often reported to come at the cost of reducing protein solubility.[17,28] For instance, stabilization strategies frequently suggest surface mutations that increase hydrophobicity, and while such muta-

[a] *Dr. S. Mazurenko*
*Loschmidt Laboratories*
*Department of Experimental Biology and RECETOX*
*Faculty of Science*
*Masaryk University*
*Zerotinovo nam. 617/9*
*601 77 Brno (Czech Republic)*
*E-mail: mazurenko@mail.muni.cz*

*This publication is part of a Special Collection on “Data Science in Catalysis” Please check the ChemCatChem  homepage for more articles in the collection.*

tions do often increase stability,[35] they tend to have a detrimental effect on solubility. The flexibilities of side chains and whole protein regions have been reported to guide the engineering of both stability[36] and solubility.[37] Moreover, the structure of mutational datasets for these two tasks is quite similar. The joint focus on both problems is, therefore, expected to bring benefit to the communities working on either task.

## 2. Data for training protein stability predictors

Recent developments in X-ray crystallography, NMR, cryo-electron microscopy allow solving protein structures at Angstrom and even sub Angstrom resolution[38] revealing the structural basis of protein binding, catalysis, and stability at the level of individual amino acids. However, such experiments are expensive, low throughput, require sophisticated instrumentation, and are often limited by the protein size. Therefore, most data on protein stability changes upon mutation come from less demanding techniques, namely differential scanning calorimetry, light scattering, circular dichroism, fluorescence spectroscopy, etc.[39] In those experiments, protein in solution is denatured by physical (temperature, pressure), chemical (pH, osmolytes), or biological (proteases) perturbation, and the output signal is recorded and analyzed. For temperature denaturation, this analysis typically yields the melting temperature $T_m$, loosely defined as the apparent midpoint of the transition in the signal, the difference in Gibbs free energy of the unfolded and folded states $\triangle G$, typically derived from data fitting, or the activity-related temperature $T_{50}$ at which the residual activity is reduced by 50% after incubation.

The pioneering effort in collecting mutational stability data from literature resulted in ProTherm,[40–42] a comprehensive database comprising numerical data from protein denaturation experiments, structural information, description of experimental methods and conditions. The overwhelming majority of protein stability change predictors were trained on the data from this database.[1] In Table 1, we summarize the most commonly used derivatives of ProTherm as well as recent additions. Unfortunately, the database was last updated in 2013 and has not been actively maintained since then. This resulted in many outdated, imprecise, or erroneous entries, which necessitated substantial manual data cleaning. Among major issues unidentified by the teams working on stability predictions[43–46] were nonmatching protein sequences and PDB entries, wrong signs and units of reported values, data incompatibility due to a wide range of experimental conditions, lack of representation for some substitutions, inadequate disclosure in the source papers. Many reported $\triangle T_m$ and $\triangle\triangle G$ values were determined under the assumption of a simple one-step reversible denaturation, whereas many proteins undergo multi-step denaturation that is not evident without proper data analysis.[47,48] The occasional presence of heat capacity difference of unfolding $\triangle C_p$ introduces a temperature dependence to $\triangle G$,[49] rendering the latter values accurate only in a narrow temperature range of the transition. However, the values reported were sometimes extrapolated to the room temperature or $T_m$ of the wild type.

Stanislav Mazurenko received his PhD. in applied mathematics and cybernetics from Lomonosov Moscow State University in 2013. He then joined the protein engineering group Loschmidt Laboratories at Masaryk University as a postdoc to work on data analysis and modelling of protein thermal denaturation. In 2018, he completed a one-year stay at the University of Liverpool, working in nonlinear optimization. He now leads a team in Loschmidt Laboratories, focusing on machine learning methods for protein engineering.

Several tendencies can be identified based on Table 1. All the datasets are restricted to single-point mutants, and in most of them only those with available PDB structure are preserved. Multiple values are averaged, and extreme conditions are sometimes excluded, as well as extreme values due to higher expected measurement errors and more significant changes to the structure of wild types upon introducing mutations. Only several teams performed a manual cleanup of the data and revealed massive inconsistencies in reported values, parameters, and structures. Moreover, data preprocessing in general varies significantly. This supports our hypothesis that the limited data size and quality might be the reason for a modest performance of ML-based predictors in independent tests.

Regarding the data structure, the wild type proteins are uniformly distributed among the four major SCOP structural classifications, as observed in S1948.[50] The authors of S1564[43] identified that the largest numbers of variants are for lysozyme (16%), followed by barnase (8%) and gene V protein (7%); most common are substitutions for alanine (26%) and substitutions from valine (11%); the least frequent are substitutions from tryptophan (only 18 out of 1564), and some substitutions are not represented at all. In each dataset, most mutations are destabilizing, and this imbalance may affect negatively the performance of the resulting predictor. Indeed, many predictors were reported to demonstrate a similar bias: mutations are usually correctly predicted as destabilizing, but those predicted stabilizing on average turn out to be neutral during experimental validation.[28]

Apart from ProTherm data, some predictors were tested on 42 mutations of the DNA binding domain of the tumor suppressor protein p53,[60] and the performance of several predictors was recently evaluated on two newly collected datasets: 96 single-point mutants of guanylate kinase[61] and 51 mutants of β-glucosidase.[33] Several teams performed an additional independent literature search, revealing the promising prospects of seeing improved protein stability predictors in the near future. Many data sets from Table 1 can also be found in VariBench – a platform for sharing published variation data for benchmarking.[58,59] Augmenting the datasets with reverse mutations with opposite signs of $\Delta\Delta G$ or $\Delta T_m$ has also gained attention recently to promote the so-called anti-symmetry of predictors: reverse mutations should produce the same predictions but with opposite signs, which turns out not to be the case for many predictors.[31,32] To provide the community with

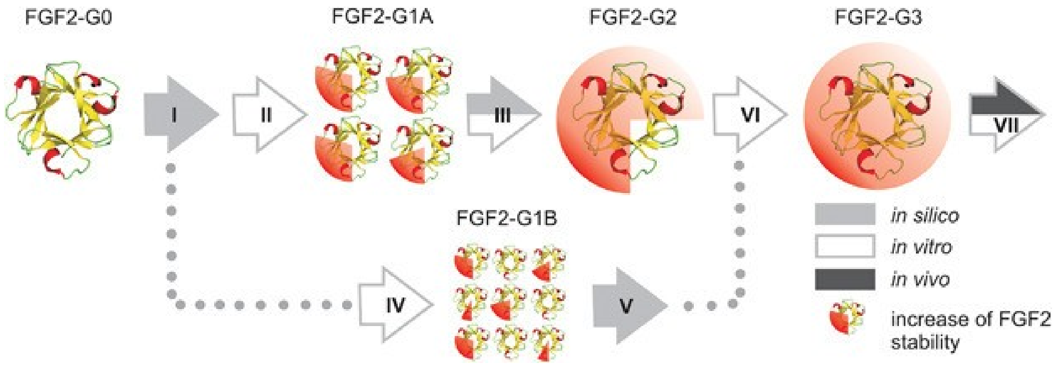
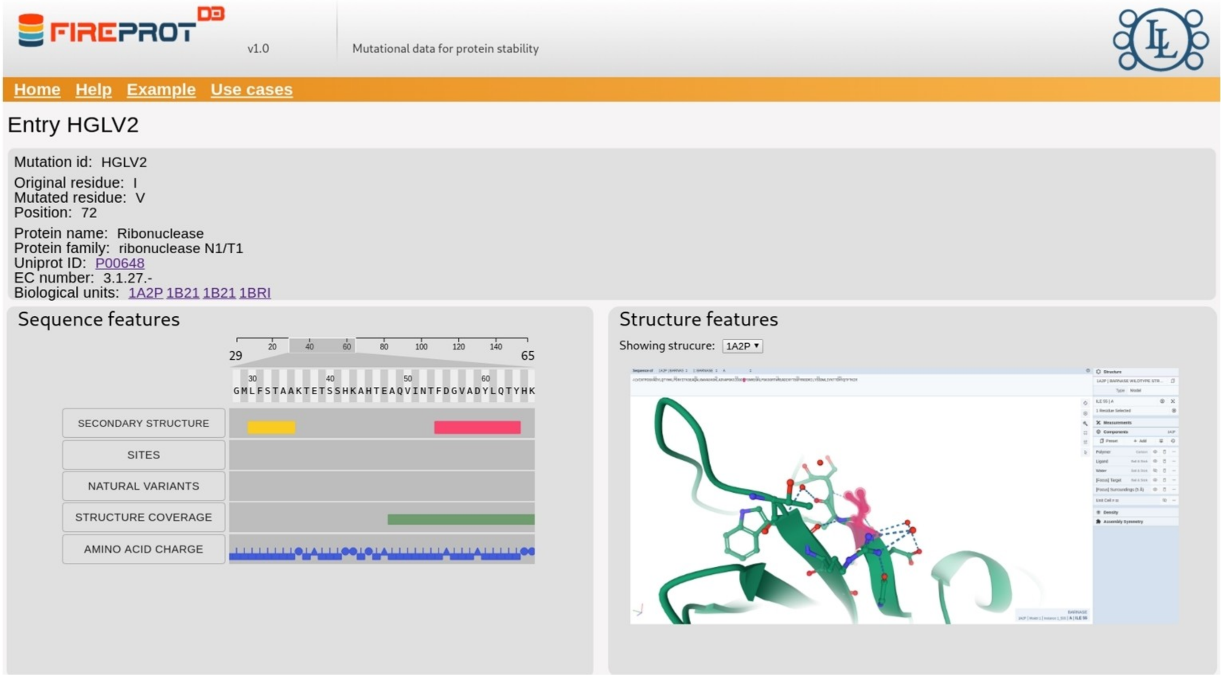


**Figure 1.** The integrated strategy combining computational analyzes with focused directed evolution for engineering the hyperstable fibroblast growth factor FGF2. Several *in silico* and *in vitro* steps resulted in the third generation protein FGF-G3 with $\triangle T_m$ of 19 °C. Reproduced with permission from Dvorak and co-workers.[15] Copyright 2018, John Wiley and Sons.

additional high quality data, we have manually processed the data from ProTherm as well as new data from literature and are depositing them to our database FireProt$^{DB}$, where they can be accessed via a user-friendly graphical interface (Figure 2). We expect to release the databased in the next few months, and its landing page can be found at loschmidt.chemi.muni.cz/fireprotdb.

## 3. Data for training protein solubility predictors

Mutational datasets for protein solubility are much more scant and heterogeneous. Protein solubility is typically defined as the concentration of folded protein in a saturated solution when in equilibrium with the solid phase. This quantity is usually estimated *in vitro* by increasing protein concentration, e.g. by adding lyophilized protein to the solvent or by protein ultrafiltration with subsequent estimating of protein fractions in the supernatant and the pellet, sometimes with the aid of various precipitants such as salts, organic solvents, or long-chain polymers.[62] At the same time, solubility can be defined more generally as *in vivo* expression, which is usually estimated as expression yield or its proxy, e.g. fluorescence intensity in split-GFP systems[63] or luminescence in split-NaNoLuc assays.[64] Protein expressibility depends on many factors as many components of a cell are involved in its synthesis and folding


FIREPROT DB v1.0 Mutational data for protein stability

Home Help Example Use cases

Entry HGLV2

Mutation id: HGLV2
Original residue: I
Mutated residue: V
Position: 72
Protein name: Ribonuclease
Protein family: ribonuclease N1/T1
Uniprot ID: P00648
EC number: 3.1.27.-
Biological units: 1A2P 1B21 1B21 1BRI

Sequence features

Structure features

Showing strucure: 1A2P ▾

**Figure 2.** The graphical user interface of FireProt$^{DB}$ (loschmidt.chemi.muni.cz/fireprotdb) containing manually curated thermostability data. The analogous web interface will also be used for the database collecting solubility data SoluProtMut$^{DB}$ (loschmidt.chemi.muni.cz/soluprotmutdb).

**Table 1.** Mutational datasets for training protein stability predictors derived from the ProTherm database.

| Name | Data points[a] | # Proteins | Comments | Availability[b] | Year |
|---|---|---|---|---|---|
| S1948[50]<br>+<br>S1925[51]<br>subset | 1948/1925 in total<br>562/553 stab.<br>30/29 neut.<br>1356/1343 dest. | 58/55 | Only single-point mutations accompanied by experimental pH, temperature, and structures at atomic resolution were considered. In S1925, 12 mutants for two proteins whose structures had missing residues, one trivial mutation, and 10 mutations with fewer than six nearest neighbors were removed. | Dataset 6 | 2005<br>2008 |
| S2648<br>+<br>S350[52]<br>subset | 2648/350 in total:<br>568/89 stab.<br>34/5 neut.<br>2046/256 dest. | 131/67 | Only single-point mutations in globular proteins with available X-ray or NMR structure were considered. Mutations in pseudo wild types and hemeproteins, those destabilizing the structure by more than 5 kcal/mol, and those involving proline were removed due to significant expected structural modifications. Multiple $\Delta\Delta G$ values were weighted-averaged, preferring pH close to 7, a temperature close to 25 °C, and no additives. The subset S350 was generated randomly to provide a benchmark. | Dataset 10 and pdf table | 2009 |
| S1109[53] | 1109 in total:<br>202 stab.<br>23 neut.<br>884 dest. | 60 | Only single-point mutations in proteins with available PDB structures were considered. The authors used Profix to fix structural defects (missing atoms, residues, or gaps), and TINKER for energy minimization, removing the proteins that failed to be processed by either tool. Only the data measured for pH 6–8 were considered assuming ionizable residues will have default charged states then. Multiple $\Delta\Delta G$ values were averaged. | Dataset 11 and xls tables | 2012 |
| S1914[45,54] | 1914 in total:<br>467 stab.<br>31 neut.<br>1416 dest. | 95 | A manually corrected subset that contains only single-point mutants. Multiple $\Delta\Delta G$ values for the same experimental conditions were averaged; for different experimental conditions, only the value closest to pH 7 was kept. Seventy four clusters of proteins with more than 25 % sequence similarity using BLASTCLUST were identified. For several measurements of the same amino acid substitution within a single cluster, only the measurement closest to pH 7 was kept. | csv file on the web server[c] | 2014 |
| S798[55] | 798 in total:<br>184 stab.<br>20 neut.<br>594 dest. | 51 | Only single-point mutations in proteins with available PDB structures were considered. Only the data measured for pH 6–8 were kept. Mutations were selected only in small and medium monomeric proteins with no more than 300 residues. Multiple $\Delta\Delta G$ values for the same experimental conditions were averaged. Mutations with the absolute $\Delta\Delta G > 10$ kcal/mol were removed due to higher expected errors. | xlsx table | 2015 |
| S1626[46] | 1626 in total:<br>463 stab.<br>58 neut.<br>1105 dest. | 90 | A manually corrected subset of $\triangle T_m$ data from ProTherm and literature search. Only single-point mutations in proteins with X-ray structures with the resolution of $< 2.5$ Å were considered. Only mutations characterized in monomeric proteins undergoing a two-state unfolding transition were included. Those with absolute $\Delta T_m$ of more than 20 °C were removed due to significant expected structural modifications. Multiple $\Delta T_m$ values for the same experimental conditions were averaged; for different experimental conditions, only the value closest to pH 7 and with the lowest concentration of additives was kept. Other thermodynamic quantities associated with mutations are also reported when available. | Dataset 14 and xls table | 2016 |
| Q306[56] | 306 in total:<br>96 stab.<br>1 neut.<br>209 dest. | 32 | Only single-point mutants with PDB IDs for wild type proteins that have a sequence identity $< 60$ % to any proteins in S2648 were considered. | Dataset 16 and csv file | 2016 |
| sDB<br>+<br>tDB[57]<br>subset | 1262/983 in total:<br>245/162 stab.<br>17/17 neut.<br>1000/804 dest. | 49/42 | Only single-point mutations in proteins with available PDB structures were considered. Only the data measured for pH 5–9 were kept. Multiple $\Delta\Delta G$ values with the variation $< 0.1$ kcal/mol were averaged. The tDB subset consists of cases with X-ray structures with no ligands. | Dataset 15 | 2016 |
| S1564[43] | 1564 in total:<br>233 stab.<br>467 neut.<br>864 dest. | 99 | A manually corrected subset of single-point mutants. The cases with $\triangle G$ values between $-0.5$ and 0.5 kcal/mol are considered neutral. No chain IDs are given, and 131 entries for 20 proteins lack PDB IDs. | Dataset 5 | 2018 |
| DeepDDG[44] | 5720/276 in total:<br>1235/77 stab.<br>77/2 neut.<br>4408/197 dest. | 242/37 | A manually corrected subset of ProTherm and literature search. Only single-point mutations in proteins with available structures were considered. The smaller subset consists of proteins with sequence identity $\leq 25$ % to proteins in S2648 and the larger set. | xlsx table | 2019 |

[a] Stab. – stabilizing, neut. – neutral (zero change if not indicated otherwise), dest. – destabilizing. [b] Dataset numbers are given as per structure.bmc.lu.se/VariBench/stability.php.[58,59] [c] The webserver is located at www.ict.griffith.edu.au/bioinf/ease.

pathways; and any perturbation of those pathways affects the solubility.

Early attempts to collect solubility data systematically at the scales suitable for general ML were made towards full sequences. In 2009, a collaborative effort of the Targeted Proteins Research Project resulted in eSoL database that comprises solubility data of around 4000 *Escherichia coli* proteins measured using the PURE cell-free expression system.[65] The more prolonged Protein Structure Initiative resulted in the TargetTrack database with more than 300 000 protein expressed and annotated.[66] Although aimed at a large-scale structure determination, it provides a proxy for quantification of protein solubility based on expressibility. The major limitation of the two databases in our context is the absence of mutational data. While some studies demonstrated potential in predicting mutational effects on solubility after training on wild type sequences only,[67] the major effort in the area was focused on assembling mutational data from existing literature (Table 2), similar to the datasets used for training protein stability predictors, and training an ML-based predictor on those datasets even despite the modest data size.

Regarding their structure, these datasets show only slight imbalance, except for CamSol dataset with just three mutations decreasing solubility. They were compiled from multiple independent publications, and the different scales for classifying solubility changes reveal that considerable effort is required to make the values compatible. In the largest data set PonSol, the number of mutants per protein ranges from 52 for Interleukin-1β to below 3 for a dozen proteins. The most common are substitutions for alanine (16%) and substitutions from leucine (11%) and lysine (10%); approximately half of the possible substitution pairs are not represented at all. A significant overlap in the data among different sets can be observed, which hinders a proper comparison of ML predictors trained on different sets. This indicates that the community will benefit significantly from a curated database resolving the overlaps as well as absorbing data published more recently. To address this limitation, we are currently working on the manually curated database SoluProtMut$^{DB}$ (loschmidt.chemi.muni.cz/soluprotmutdb) that will comprise both the data systematically collected from published sources and the experimental data collected in our laboratory.

## 4. Perspectives

The analysis of the literature presented in this study demonstrates how challenging the task of collecting mutational data is even for such habitually measured protein properties as stability and solubility. Apart from data scarcity, which is arguably most urgent in the latter case, the data quality requires much attention. This problem comes in different flavors: from inaccuracies, insufficient disclosure, and lack of standard protocols of data analysis in the original publications, to the errors and difficulties of aggregating information from different sources, biases and imbalances in the resulting datasets. Therefore, the community of researchers developing ML predictors of protein stability and solubility changes will greatly benefit from up-to-date, manually curated, user-friendly, and ML-friendly databases.

Manual curation is indispensable, as demonstrated by the teams that had to discard or change the majority of data from ProTherm (Table 1) due to erroneous values, incorrect or missing structures and sequences, non-existent substitutions, and ambiguous experimental conditions. Many other derivative datasets were not cleaned thoroughly, compromising the quality of the resulting predictors. However, this does not come as a surprise, since as an ML developer, one might have neither enough resources nor proper expertise to check the sources and validate the quality of experiments for each data point. Moreover, with the lack of an updatable database to report inconsistencies and compare dataset overlaps, one has to repeat the cleaning steps almost from scratch every time before training a predictor on more recent data. This repetition leads to a waste of time and delays the maturation of the field into the next stage of ML development, e.g. in-depth analysis and interpretation of successful predictors to uncover biophysical mechanisms behind better predictions.

User-friendliness in terms of graphical summary and statistics will allow faster monitoring of the structure of the data to reveal biases in real-time. These refer to over-represented proteins or protein families, amino acids chosen for

**Table 2.** Mutational datasets for training predictors of protein solubility change.

| Name | Data points | # Proteins | Comments | Availability[a] | Year |
|---|---|---|---|---|---|
| OptSolMut[68] | 137 in total:<br>59 increased<br>78 decreased | 19 | Binary classification for single- and multiple-point mutants from 15 published studies, with PDB IDs provided. Among 105 single-point mutants, 61 decreased and 44 increased solubility compared to wild types. In total, 121 mutants were soluble both before and after the mutation, but the extent of solubility changed. Also 26 mutants have stability changes reported. | xls table | 2010 |
| CamSol[69] | 63 in total:<br>53 increased<br>7 neutral<br>3 decreased | 19 | Binary classification for single- and multiple-point mutants from 4 published studies, with sequences of wild types and mutants provided. Among 40 single-point mutants, 1 decreased and 38 increased solubility compared to wild types. | xls table | 2014 |
| Aggrescan3D[70] | 129 in total:<br>87 increased<br>42 decreased | 29 | Binary classification for single- and multiple-point mutants from 28 published studies, with PDB IDs provided. Among 106 single-point mutants, 37 decreased and 69 increased solubility compared to wild types. | pdf table | 2015 |
| PonSol[71]<br>+<br>SODA[72]<br>subset | 443 in total:<br>85 increased<br>222 neutral<br>136 decreased | 71 | Five-level classification for single-point mutants from >80 published studies using a multi-step literature search and data mining tools. Mutations affecting aggregation were excluded due to a different physicochemical phenomenon. Links to wild type sequences are provided for all proteins, and PDB IDs for 14 proteins. Among 136 mutations decreasing solubility, 46 were classified as decreasing significantly. Among 85 mutations increasing solubility, 27 were classified as increasing significantly. | xlsx table | 2016 |
| | 145 in total:[b]<br>61 increased<br>84 decreased | 49 | All neutral mutations, as well as ambiguous examples such as those with sequence mismatches, were deleted from PonSol dataset. Only proteins with a maximum pairwise sequence identity of <30% with the CamSol dataset were kept. Links to wild type sequences are provided. | csv table | 2017 |

[a] Data with their assignment to individual datasets will be available in the database SoluProtMut$^{DB}$: loschmidt.chemi.muni.cz/soluprotmutdb. [b] 142 reported in the original paper and 145 available at http://protein.bio.unipd.it/soda/about.

mutation or those substituting, locations of the mutations, e.g. with respect to the sequence, secondary structure elements, protein surface, tunnels, active sites, etc. The identification of such biases is of critical importance in ML – a data-driven strategy unable to correct data biases automatically, without additional tweaking. The prediction power of an ML-based model has yet to be explored for the poorly represented substitutions or proteins with low homology or different unfolding patterns than those in the training data.

The demand from the ML side also comes for the structure of such a database. The precise identification of mutations, corresponding sequences, and PDB IDs is one ingredient. Another one is adhering to the tidy data principles,[73] i.e. data representation in a clear table format where columns correspond to variables, such as substitutions, protein identifiers, experimental conditions, etc., and each row corresponds to experimental observation. While these principles seem easy to implement, representing multiple-point mutations or new experimental setups will challenge the database developers.

An interesting recent initiative is ProtaBank[74] – the database aimed to collect protein engineering data in one place, including some of the datasets mentioned earlier. The creators opted to target a wide range of assays, an excellent idea given the increasing interest in data generation and lack of any central repository of this kind. They also offered several search tools to analyze comprehensively published results concerning a particular sequence inquiry, including related sequences given by BLAST search. On the other hand, the wide focus and variability of the supported data types come at the cost of increasing the effort required for fetching all the available data, e.g. protein stability or solubility changes, and processing them into ML-friendly format. With this in mind, we are currently working on two manually curated ML-oriented mutational databases to be officially released in 2020: FireProt$^{DB}$ for protein stability (loschmidt.chemi.muni.cz/fireprotdb) and SoluProtMut$^{DB}$ (loschmidt.chemi.muni.cz/soluprotmutdb) for protein solubility changes. The preliminary versions include ca. 14 000 single-point mutants in around 270 proteins and over 10 000 data points from 100 proteins, respectively. Interestingly, most of the teams, including ours, have resorted to manual search for data in literature so far. Thus, automated data mining remains largely unexplored in this respect.[71,75] The mutational datasets discussed earlier present significant challenges in this respect since the information about mutations and their effects is usually scattered across the publication, and additional effort is required, e.g., to identify automatically whether a positive value of △△G found in a text means increased or decreased stability.

Regarding the perspectives in generating new data, several recent experimental techniques raise hopes of significantly enhancing the available data on mutational changes. In particular, deep mutational scanning[76] that couples next-generation sequencing[77,78] with high-throughput assays, e.g. based on fluorescence-activated cell sorting.[79,80] This approach links genotype to phenotype by synthesizing a large library of mutant sequences, selecting for expressed phenotypes, and sequencing the library before and after the selection to quantify the fitness of each mutant. The screening protocols are being actively developed to represent fitness from various angles, and some of them already approximate protein stability and solubility.[81–83] Two major advantages of this approach are the data size and distribution. Data sets generated by deep mutational scanning can easily run into thousands or tens of thousands of mutants, which is terrific news for data-hungry ML. The library generated often covers the space of possible mutations quite uniformly, which compares favorably with more standard low-throughput approaches, in which the selection of variants is usually skewed towards anticipated best performers and negative results are sometimes discarded. Therefore, we expect many new exciting data sets in the near future, which is likely to open up new opportunities for using more powerful ML architectures such as artificial neural networks. Several recent reviews identified the trend in biocatalyst design towards using nonlinear ML models compared to predominantly simple linear predictors in the past.[25,84] And such a transition will lead to more accurate and generalizable tools once a sufficient amount of data is available to steer the flexibility granted by the nonlinear models.

It is also most desirable if the newly collected data are published according to the FAIR principles[85,86] that are created to encourage authors to take data sharing, discoverability, and reuse into account from the outset of preparing their results. These principles stipulate that data should be identified, described, and indexed clearly and unequivocally, should use standard technical and semantic data formats, variables, and ontologies, and should provide clearly defined access procedures, ideally by automated means. Regarding the application of those principles to publications with mutational data, the following guidelines will help promote the collection of high-quality data sets for training and validating predictors:

- Include and examine protein sequence identifiers, PDB ID's, annotations of mutations, etc. in publications. Any inaccuracies in reporting propagate into databases, require significantly more effort in identifying at later stages, and often lead to discarding the data. This is an undesirable outcome for all the parties involved since the data are not reused and their scientific impact is curtailed.
- Report and upload as much data as possible, even for those mutations that did not lead to the desired outcome. Detailed numerical data is often provided only for several mutants out of all those tested, and the rest are reported in an aggregated format only, such as on a graph or in a table of descriptive statistics, precluding their usage in ML training.
- Publish the data as supplements to the original publication, where they are less likely to be lost. Personal data storages get closed, group pages move to new locations, and departments get restructured, which is why many datasets are now unavailable as their links stopped working.[1]
- Add a csv or xls data table, preferably already in the "tidy" format, even when some of the values are already reported in the main text to improve the access to your data, promote your work among the bioinformatics community, and increase its impact.

Finally, we would urge companies to release their data sets, which might be difficult for their ongoing projects due to undesirable disclosure but should be feasible for past results. The power of ML-based predictors comes from exploiting all the available data, and while individual gains are not always apparent, the whole protein engineering community will benefit from time, effort, and resources saved using predictors that are more accurate. Data scarcity is now the major bottleneck for developing more precise predictors, and if we want to accelerate the research of human neurodegenerative disorders, metabolic diseases, cancer, produce more efficient drugs, and widen the industrial application of biocatalysts, sharing your data is a small piece of the puzzle that might lead to bigger improvements.

## Acknowledgements

*This work was supported by grants from the Czech ministry of Education, Youth and Sports (LM2015047, LM2018121, 02.1.01/0.0/0.0/18_046/0015975) and the European Union (CETOCOEN EXCELLENCE Teaming 2 project supported by Horizon2020 857560). SM was supported by the Operational Programme Research, Development and Education project MSCAfellow@MUNI (CZ.02.2.69/0.0/0.0/17_050/0008496).*

## Conflict of Interest

The authors declare no conflict of interest.

**Keywords:** Database · Machine learning · Protein design · Protein engineering · Protein modifications

[1] M. Musil, H. Konegger, J. Hon, D. Bednar, J. Damborsky, *ACS Catal.* **2019**, *9*, 1033–1054.
[2] Q. Liu, G. Xun, Y. Feng, *Biotechnol. Adv.* **2019**, *37*, 530–537.
[3] C. J. Roberts, *Curr. Opin. Biotechnol.* **2014**, *30*, 211–217.
[4] D. Esposito, D. K. Chatterjee, *Curr. Opin. Biotechnol.* **2006**, *17*, 353–358.
[5] H. Hussain, D. I. Fisher, R. G. Roth, W. Mark Abbott, M. A. Carballo-Amador, J. Warwicker, A. J. Dickson, *FEBS Lett.* **2018**, *592*, 2499–2511.
[6] C. M. Dobson, *Nature* **2003**, *426*, 884–890.
[7] P. Ciryam, G. G. Tartaglia, R. I. Morimoto, C. M. Dobson, M. Vendruscolo, *Cell Rep.* **2013**, *5*, 781–790.
[8] F. Chiti, C. M. Dobson, *Annu. Rev. Biochem.* **2017**, *86*, 27–68.
[9] J. L. Silva, C. V. D. M. Gallo, D. C. F. Costa, L. P. Rangel, *Trends Biochem. Sci.* **2014**, *39*, 260–267.
[10] U. P. Andley, M. A. Reilly, *Exp. Eye Res.* **2010**, *90*, 699–702.
[11] A. Meulemans, S. Seneca, T. Pribyl, J. Smet, V. Alderweirldt, A. Waeytens, W. Lissens, R. Van Coster, L. De Meirleir, J. P. Di Rago, et al., *J. Biol. Chem.* **2010**, *285*, 4099–4109.
[12] C. C. Lee, J. M. Perchiacca, P. M. Tessier, *Trends Biotechnol.* **2013**, *31*, 612–620.
[13] D. Bednar, K. Beerens, E. Sebestova, J. Bendl, S. Khare, R. Chaloupkova, Z. Prokop, J. Brezovsky, D. Baker, J. Damborsky, *PLoS Comput. Biol.* **2015**, *11*, e1004556.
[14] M. Musil, J. Stourac, J. Bendl, J. Brezovsky, Z. Prokop, J. Zendulka, T. Martinek, D. Bednar, J. Damborsky, *Nucleic Acids Res.* **2017**, *45*, 393–399.
[15] P. P. Dvorak, D. Bednar, P. Vanacek, L. Balek, L. Eiselleova, V. Stepankova, E. Sebestova, M. Kunova Bosakova, Z. Konecna, S. Mazurenko, et al., *Biotechnol. Bioeng.* **2018**, *115*, 850–862.
[16] K. Trainor, A. Broom, E. M. Meiering, *Curr. Opin. Struct. Biol.* **2017**, *42*, 136–146.
[17] A. Broom, Z. Jacobi, K. Trainor, E. M. Meiering, *J. Biol. Chem.* **2017**, *292*, 14349–14361.
[18] C. Silva, M. Martins, S. Jing, J. Fu, A. Cavaco-Paulo, *Crit. Rev. Biotechnol.* **2018**, *38*, 335–350.
[19] R. Kazlauskas, *Chem. Soc. Rev.* **2018**, *47*, 9026–9045.
[20] J. S. Ebo, N. Guthertz, S. E. Radford, D. J. Brockwell, *Curr. Opin. Struct. Biol.* **2020**, *60*, 157–166.
[21] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, *Nucleic Acids Res.* **2008**, *36*, D202–D205.
[22] E. H. Kellogg, A. Leaver-Fay, D. Baker, *Proteins Struct. Funct. Bioinf.* **2011**, *79*, 830–838.
[23] F. Cadet, N. Fontaine, G. Li, J. Sanchis, M. Ng, F. Chong, R. Pandjaitan, I. Vetrivel, B. Offmann, M. T. Reetz, *Sci. Rep.* **2018**, *8*, 16757.
[24] G. Li, Y. Dong, M. T. Reetz, *Adv. Synth. Catal.* **2019**, *361*, 2377–2386.
[25] S. Mazurenko, Z. Prokop, J. Damborsky, *ACS Catal.* **2020**, *10*, 1210–1223.
[26] P. Domingos, *Commun. ACM* **2012**, *55*, 78–87.
[27] V. Marx, *Nat. Methods* **2019**, *16*, 463–467.
[28] A. Broom, K. Trainor, Z. Jacobi, E. M. Meiering, *Structure* **2020**, *28*, 1–10.
[29] S. Khan, M. Vihinen, *Hum. Mutat.* **2010**, *31*, 675–684.
[30] V. Potapov, M. Cohen, G. Schreiber, *Protein Eng. Des. Sel.* **2009**, *22*, 553–560.
[31] F. Pucci, K. V. Bernaerts, J. M. Kwasigroch, M. Rooman, *Bioinformatics* **2018**, *34*, 3659–3665.
[32] J. Fang, *Briefings Bioinf.* **2019**, *2019*, 1–8.
[33] P. Huang, S. K. S. Chu, H. N. Frizzo, M. P. Connolly, R. W. Caster, J. B. Siegel, *ACS Omega* **2020**, *5*, 6487–6493.
[34] A. Ciechanover, A. Orian, A. L. Schwartz, *BioEssays* **2000**, *22*, 442–451.
[35] A. Nisthal, C. Y. Wang, M. L. Ary, S. L. Mayo, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 16367–16377.
[36] Z. Sun, Q. Liu, G. Qu, Y. Feng, M. T. Reetz, *Chem. Rev.* **2019**, *119*, 1626–1665.
[37] B. K. Bhandari, P. P. Gardner, C. S. Lim, *Bioinformatics* **2020**, btaa578.
[38] U. K. Eriksson, G. Fischer, R. Friemann, G. Enkavi, E. Tajkhorshid, R. Neutze, *Science* **2013**, *340*, 1346–1349.
[39] S. Mazurenko, J. Stourac, A. Kunka, S. Nedeljković, D. Bednar, Z. Prokop, J. Damborsky, S. Nedeljkoví, D. Bednar, Z. Prokop, et al., *Nucleic Acids Res.* **2018**, *46*, W344–W349.
[40] M. D. S. Kumar, K. A. Bava, M. M. Gromiha, P. Prabakaran, K. Kitajima, H. Uedaira, A. Sarai, *Nucleic Acids Res.* **2006**, *34*, D204–D206.
[41] K. A. Bava, M. M. Gromiha, H. Uedaira, K. Kitajima, A. Sarai, *Nucleic Acids Res.* **2004**, *32*, D120–D121.
[42] M. M. Gromiha, J. An, H. Kono, A. Sarai, M. Oobatake, H. Uedaira, A. Sarai, *Nucleic Acids Res.* **1999**, *27*, 286–288.
[43] Y. Yang, S. Urolagin, A. Niroula, X. Ding, B. Shen, M. Vihinen, *Int. J. Mol. Sci.* **2018**, *19*, 1009.
[44] H. Cao, J. Wang, L. He, Y. Qi, J. Z. Zhang, *J. Chem. Inf. Model.* **2019**, *59*, 1508–1514.
[45] L. Folkman, B. Stantic, A. Sattar, *BMC Genomics* **2014**, *15*, S6.
[46] F. Pucci, R. Bourgeas, M. Rooman, *J. Phys. Chem. Ref. Data* **2016**, *45*, 023104.
[47] M. Tsytlonok, L. S. Itzhaki, *Arch. Biochem. Biophys.* **2013**, *531*, 14–23.
[48] S. Mazurenko, A. Kunka, K. Beerens, C. M. Johnson, J. Damborsky, Z. Prokop, *Sci. Rep.* **2017**, *7*, 1–14.
[49] P. L. Privalov, A. I. Dragan, *Biophys. Chem.* **2007**, *126*, 16–24.
[50] E. Capriotti, P. Fariselli, R. Casadio, *Nucleic Acids Res.* **2005**, *33*, W306–W310.
[51] M. Masso, I. I. Vaisman, *Bioinforma. Orig. Pap.* **2008**, *24*, 2002–2009.
[52] Y. Dehouck, A. Grosfils, B. Folch, D. Gilis, P. Bogaerts, M. Rooman, *Bioinformatics* **2009**, *25*, 2537–2543.
[53] Z. Zhang, L. Wang, Y. Gao, J. Zhang, M. Zhenirovskyy, E. Alexov, *Bioinformatics* **2012**, *28*, 664–671.
[54] L. Folkman, B. Stantic, A. Sattar, Y. Zhou, *J. Mol. Biol.* **2016**, *428*, 1394–1405.
[55] L. Jia, R. Yarlagadda, C. C. Reed, *PLoS One* **2015**, *10*, e0138022.
[56] L. Quan, Q. Lv, Y. Zhang, *Bioinformatics* **2016**, *32*, 2936–2946.
[57] I. Getov, M. Petukh, E. Alexov, *Int. J. Mol. Sci.* **2016**, *17*, 512.
[58] P. Sasidharan Nair, M. Vihinen, *Hum. Mutat.* **2013**, *34*, 42–49.
[59] A. Sarkar, Y. Yang, M. Vihinen, *Database* **2020**, *2020*, 117.
[60] D. E. V. Pires, D. B. Ascher, T. L. Blundell, *Bioinformatics* **2014**, *30*, 335–342.
[61] K. N. McGuinness, W. Pan, R. P. Sheridan, G. Murphy, A. Crespo, *PLoS One* **2018**, *13*, e0203819.
[62] R. M. Kramer, V. R. Shende, N. Motl, C. N. Pace, J. M. Scholtz, *Biophys. J.* **2012**, *102*, 1907–1915.
[63] M. A. Lockard, P. Listwan, J.-D. Pedelacq, S. Cabantous, H. B. Nguyen, T. C. Terwilliger, G. S. Waldo, *Protein Eng. Des. Sel.* **2011**, *24*, 565–578.

[64] A. S. Dixon, M. K. Schwinn, M. P. Hall, K. Zimmerman, P. Otto, T. H. Lubben, B. L. Butler, B. F. Binkowski, T. MacHleidt, T. A. Kirkland, et al., *ACS Chem. Biol.* **2016**, *11*, 400–408.

[65] T. Niwa, B. W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, H. Taguchi, *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 4201–4206.

[66] H. M. Berman, M. J. Gabanyi, A. Kouranov, D. I. Micallef, J. Westbrook, *Protein Structure Initiative - TargetTrack 2000–2017 - All Data Files*, **2017**.

[67] D. Raimondi, G. Orlando, P. Fariselli, Y. Moreau, D. Raimondiid, G. Orlando, P. Fariselli, Y. Moreauid, *PLoS Comput. Biol.* **2020**, *16*, e1007722.

[68] Y. Tian, C. Deutsch, B. Krishnamoorthy, *Algorithms Mol. Biol.* **2010**, *5*, 33.

[69] P. Sormanni, F. A. Aprile, M. Vendruscolo, *J. Mol. Biol.* **2015**, *427*, 478–490.

[70] R. Zambrano, M. Jamroz, A. Szczasiuk, J. Pujols, S. Kmiecik, S. Ventura, *Nucleic Acids Res.* **2015**, *43*, W306–W313.

[71] Y. Yang, A. Niroula, B. Shen, M. Vihinen, *Bioinformatics* **2016**, *32*, 2032–2034.

[72] L. Paladin, D. Piovesan, S. C. E. Tosatto, *Nucleic Acids Res.* **2017**, *45*, W236–W240.

[73] H. Wickham, *J. Stat. Softw.* **2014**, *59*, 1–23.

[74] C. Y. Wang, P. M. Chang, M. L. Ary, B. D. Allen, R. A. Chica, S. L. Mayo, B. D. Olafson, *Protein Sci.* **2018**, *27*, 1113–1124.

[75] J. R. Kitchin, *Nat. Can.* **2018**, *1*, 230–232.

[76] D. M. Fowler, S. Fields, *Nat. Methods* **2014**, *11*, 801–807.

[77] J. K. Kulski, *Next Generation Sequencing - Advances, Applications and Challenges*, InTech, **2016**.

[78] J. Straiton, T. Free, A. Sawyer, J. Martin, *BioTechniques* **2019**, *66*, 60–63.

[79] H. A. Bunzel, X. Garrabou, M. Pott, D. Hilvert, *Curr. Opin. Struct. Biol.* **2018**, *48*, 149–156.

[80] P. Jacques, M. Béchet, M. Bigan, D. Caly, G. Chataigné, F. Coutte, C. Flahaut, E. Heuson, V. Leclère, D. Lecouturier, et al., *Bioprocess Biosyst. Eng.* **2017**, *40*, 161–180.

[81] A. J. Riesselman, J. B. Ingraham, D. S. Marks, *Nat. Methods* **2018**, *15*, 816–822.

[82] K. A. Matreyek, L. M. Starita, J. J. Stephany, B. Martin, M. A. Chiasson, V. E. Gray, M. Kircher, A. Khechaduri, J. N. Dines, R. J. Hause, et al., *Nat. Genet.* **2018**, *50*, 874–882.

[83] J. R. Klesmith, J. P. Bacik, E. E. Wrenbeck, R. Michalczyk, T. A. Whitehead, *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 2265–2270.

[84] W. Yang, T. T. Fidelis, W.-H. Sun, *ACS Omega* **2020**, *5*, 83–88.

[85] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., *Sci. Data* **2016**, *3*, 1–9.

[86] M. Boeckhout, G. A. Zielhuis, A. L. Bredenoord, *Eur. J. Hum. Genet.* **2018**, *26*, 931–936.

Manuscript received: June 3, 2020
Revised manuscript received: July 8, 2020
Accepted manuscript online: July 20, 2020
Version of record online: September 3, 2020