Machine learning predicts immunoglobulin light chain toxicity through somatic mutations

Maura Garofalo[1], Luca Piccoli[1], Margherita Romeo[2], Sara Ravasio[1,3], Mathilde Foglierini[1,4], Milos Matkovic[1], Jacopo Sgrignani[1], Marco Prunotto[5], Luisa Diomede[2], Olivier Michielin[6,7], Antonio Lanzavecchia[1], and Andrea Cavalli[1,4*]

[1]Institute for Research in Biomedicine, Università della Svizzera italiana, Bellinzona, Switzerland

[2]Department of Molecular Biochemistry and Pharmacology, Istituto di Ricerche Farmacologiche Mario negri IRCCS, Milan, Italy

[3]Institute of Microbiology, ETH Zurich, Zurich, Switzerland

[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[5]School of Pharmaceutical Sciences, Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland

[6]Molecular Modeling Group, SIB Swiss Institute of Bioinformatics, University of Lausanne, Quartier UNIL-Sorge, Bâtiment Amphipôle, Lausanne, Switzerland

[7]Department of Oncology, University Hospital of Lausanne, Ludwig Cancer Research - Lausanne Branch, Lausanne, Switzerland

Corresponding author

Andrea Cavalli

Institute for Research in Biomedicine

Università della Svizzera italiana

Via Vincenzo Vela 6

CH-6500 Bellinzona, Switzerland

Email: andrea.cavalli@irb.usi.ch

Abstract

In systemic light chain amyloidosis (AL), pathogenic monoclonal immunoglobulin light chains (LCs) form toxic aggregates and amyloid fibrils in target organs. Prompt diagnosis is crucial to avoid permanent organ damage. However, delays in diagnosis are common, with a consequent poor patient's prognosis, as symptoms usually appear only after strong organ involvement. Here, we present LICTOR, a machine learning approach predicting LC toxicity in AL, based on the distribution of somatic mutations acquired during clonal selection. LICTOR achieved a specificity and a sensitivity of 0.82 and 0.76, respectively, with an area under the receiver operating characteristic curve (AUC) of 0.87. Tested on an independent set of 12 LCs sequences with known clinical phenotypes, LICTOR achieved a prediction accuracy of 83%. Furthermore, we were able to abolish the toxic phenotype of an LC by *in silico* reverting two germline-specific somatic mutations identified by LICTOR and by experimentally assessing the loss of toxicity in a *Caenorhabditis elegans* model *in vivo*. Therefore, LICTOR represents a promising strategy for AL diagnosis and reducing high mortality rates in AL.

Introduction

Systemic light chain amyloidosis (AL) is a monoclonal gammopathy characterized by the abnormal proliferation of a plasma cell clone producing large amounts of pathogenic immunoglobulin free light chains (LCs)[1]. LCs, mainly secreted as homodimers[2], misfold forming toxic species and amyloid fibrils which accumulate in target organs and lead to fatal organ dysfunction and death[1]. Although LCs deposition can occur in any organ except the brain, the kidney and heart are the most affected sites, with the latter bearing the worst prognosis. Symptoms of AL are nonspecific and usually reflect advanced organ involvement. Therefore, an early diagnosis is essential to avoid irreversible organ damage. However, the

complexity of the disease and its vague symptoms make a timely diagnosis of AL extremely challenging[3,4].

Pre-existing monoclonal gammopathy of undetermined significance (MGUS) is a known risk factor for developing AL, with 9% of MGUS patients progressing to AL[5-7]. However, early diagnosis is still difficult since reliable diagnostic tests predicting whether MGUS patients are likely to develop AL are currently lacking[7,8]. Predicting the onset of AL is highly challenging, as each patient carries a different pathogenic LC sequence resulting from a unique rearrangement of variable (V) and joining (J) immunoglobulin genes and a unique set of somatic mutations (SMs) acquired during B cell affinity maturation[9] (Fig. 1a). Therefore, the development of a specific prediction tool represents a crucial step to anticipate AL diagnosis and improve patients' prognosis.

Machine learning techniques are becoming very prominent in different areas of science and are also gaining acceptance in medicine. Indeed, machine learning has been used in different areas of medicine, such as diagnosis[10-12], prognosis[13,14], drug discovery[15,16] and drug sensitivity prediction[17-19]. In these approaches, machines learn information from data without being explicitly programmed and simulate human intelligence to make predictions[20]. The high diversity of LC sequences accountable for AL development and the possibility of accessing databases of pathogenic and non-pathogenic LC sequences prompted us to use a machine learning-based strategy to devise a predictor of LC toxicity in AL named LICTOR (λ-**LI**ght-**C**hain **TO**xicity predicto**R**).

LICTOR uses SMs as predictor variables based on the hypothesis that SMs are the main LC toxicity-discriminating factors, achieving a specificity and a sensitivity of 0.82 and 0.76, respectively, with an AUC of 0.87. LICTOR performance was assessed with an independent set of LCs with a known clinical phenotype but not used in the training. Furthermore, to experimentally validate LICTOR, we used our predictor to abolish the pathological phenotype

of a cardiotoxic LC and verified the outcome with a *Caenorhabditis elegans*-based assay[21,22]. This assay takes advantage of the evolutionary relationship between the nematode pharynx and the vertebrate heart, evaluating the reduction of the *C. elegans* pharyngeal pumping rate after administration of cardiotoxic LCs as a measure of proteotoxicity[21,22]. Both experiments recapitulated LICTOR's predictions.

Taken together, these results confirm that LICTOR provides insights into specific features differentiating toxic and non-toxic LCs. Therefore, it may represent a powerful tool to improve AL diagnosis and unveil a novel strategy for patient treatment through personalized medicine.

Results

Somatic mutations are key LC toxicity-discriminating factors

To investigate the role of SMs in the generation of toxic LCs and validate their use as predictor variables in a LC toxicity predictor, we collected a database of 1,075 λ LC sequences. The database included 428 "toxic" sequences from AL patients (hereafter referred to as *tox*) and 647 "non-toxic" LCs (*nox*) comprising sequences from healthy donor repertoires, other autoimmune diseases or cancer, obtained from ALBase[23] (428 *tox*, 590 *nox*) and an in-house LCs database not related to AL (57 *nox)*. We restricted our analysis to λ LCs since this isotype is more prevalent than the kappa (κ) isotype in AL patients (λ/κ=3:1 compared to that of healthy individuals, λ/κ=1:2)[24]. To identify SMs, all LCs were aligned to the corresponding germline (GL) sequence obtained using the IMGT database[25]. LCs were then numbered according to the Kabat-Chothia scheme (using progressive enumeration from 1 to 125), allowing the structural comparison of LCs with different sequence lengths (Methods and Fig. 1b). Next, we counted the number of mutated (M) and non-mutated (NM) residues at each position *i* in *tox* and *nox* sequences (tox_M^i and nox_M^i, tox_{NM}^i and nox_{NM}^i, respectively) and used

Fisher's exact test[26] to assess whether the frequencies of mutations (tox^i_M, nox^i_M) and non-mutations (tox^i_{NM}, nox^i_{NM}) were significantly ($p < 0.05$) different. Finally, the odds ratio (OR)[26] was used to assess the association strength between mutations and toxicity at each position *i* in *tox* and *nox* sequences (Methods and Fig. 1c). Interestingly, 48 of 53 positions with a statistically significant difference ($p < 0.05$) between the two groups (Fig. 1c) showed a higher rate of mutation in the *tox* group (OR >1), while only 5 positions reported a higher mutation rate in the *nox* group (OR <1). To exclude possible bias induced by the use of a group of *nox* sequences having an artificially low level of SMs, we randomly selected 1,000 LC sequences from a healthy donor repertoire (*hdnox*)[27] and compared the probability distributions of the number of SMs (PDSM) between the three groups. We observed similar PDSM between the *nox* and *hdnox* groups, while the PDSM of *tox* and *hdnox*, as well as *tox* and *nox*, were significantly different. This result supports *nox* sequences as a bona fide group of LCs (Supplementary Figure 1). Overall, these findings suggest that SMs are key determinants of the toxicity of LCs and, thus, can be used as predictor variables to develop LC toxicity prediction tools.

Prediction of LC toxicity using machine learning

The previous findings prompted us to use SMs as features to develop a machine learning approach automatically classifying LCs as either toxic or non-toxic in AL. To this end, we combined the information from SMs with knowledge of the 3D structure of LCs homodimers[28,29] to create three families of predictor variables used in the training of machine learning algorithms. The first family, termed AMP (Amino acid at each Mutated Position), highlights sequence features, identifying the presence or absence of an SM at each position of the LCs sequences. The second family, termed MAP (Monomeric Amino acid Pairs), identifies the presence or absence of mutations in residues in close contact in the LC monomeric 3D

structure (distance <7.5 Å). Finally, the third family, named DAP (Dimeric Amino acid Pairs), identifies the presence or absence of mutations at positions in close contact but belonging to different chains. Next, four machine learning algorithms (Bayesian network, logistic regression, J48 and random forest)[30] were evaluated for their ability to solve the classification problem, using our database as input. To assess the importance of the different classes of predictor variables, we performed 28 prediction experiments including all possible combinations of AMP, MAP and DAP families. In addition, to avoid unbalanced class problems, i.e., the tendency of machine learning algorithms to assign sequences to the largest class in the dataset, *nox* in our case, each of the 28 experiments was performed with and without balancing the training set using a SMOTE (Synthetic Minority Over-sampling TEchnique) filter[31]. The assessment of algorithms and predictor variable combinations was performed using 10-fold cross-validation to avoid overfitting. We found that for all tested machine learning algorithms, the best combination of predictor variable families provided an area under the receiver operating characteristic curve (AUC) that substantially differed from that of a random classifier (0.50), with random forest being the best classifier (0.87) and J48 being the worst (0.75) (Fig. 2a and Supplementary Table 1). Furthermore, all four classifiers relied on the AMP family to predict LC toxicity, while only random forest used all three families of predictor variables (AMP+MAP+DAP) in its best configuration. Overall, these findings highlight the importance of the structural context of somatic mutations in defining the toxicity of an LC and identify random forest using AMP, MAP, and DAP as the best approach in our case. For this reason, we used random forest in our implementation of LICTOR.

Validation of LICTOR

Next, we sought to validate the prediction accuracy of LICTOR with a set of sequences with known clinical phenotypes but not present in the training set (*valset*).[29] The *valset* (Supplementary Table 2) comprised a total of 12 LCs, including 7 sequences associated with AL with cardiac involvement (H3, H6, H7, H9, H15, H16 and H18) and 5 from multiple myeloma patients (M2, M7, M8, M9 and M10). LICTOR was able to correctly classify 10 (6 *tox* and 4 *nox*) of 12 LCs as either toxic or non-toxic (Supplementary Table 2 and Fig. 3a). The probability of achieving a similar accuracy with a random classifier was 0.016, strengthening the argument that LICTOR is a robust and accurate tool to predict the clinical toxicity of previously unseen LCs.

To further assess the robustness of LICTOR, we performed two additional tests. In the first test, we predicted the toxicity of 100 randomly selected LCs from the healthy donor repertoire *hdnox* (all absent from the training set). In this case, LICTOR correctly classified 80% of the sequences as non-toxic (Supplementary Table 3), thus confirming a similar level of accuracy as for the training set. In the second test, we assessed LICTOR by further verifying the absence of overfitting. To achieve this, we took all *tox* sequences and randomly labelled half of them as *nox*. Then, we trained a classifier using the 10-fold cross-validation on such a dataset obtaining an AUC of 0.5 (Supplementary Table 4), equivalent to that of a random classifier. The same procedure has been used for the *nox* group (by randomly assigned half of them as *tox* and training a classifier on such a dataset), obtaining also in this case an AUC of 0.5 (Supplementary Table 5). These results further underlines that *tox* and *nox* sequences have distinctive features allowing their discrimination.

LC germline VJ rearrangement information does not improve prediction performance

To further underscore the role of SMs as main discriminants between toxic and non-toxic LCs, we trained the same machine learners employed before, using the LC germline VJ

rearrangements as a unique predictor variable, given the well-documented overrepresentation of certain VL germline genes in AL[32,33]. All the resulting *germline-based* classifiers achieved an AUC of 0.77 in their best configuration (Fig. 2b and Supplementary Table 6), a value substantially better than that of a random classifier, although much lower than LICTOR's score (0.87). Interestingly, adding LC germline VJ rearrangements to LICTOR did not improve its prediction performance (Supplementary Table 7).

Next, we computed the specificity and sensitivity of the two random forest predictors, LICTOR and the *germline-based* predictor, maximizing the Youden index (J)[34] as a function of the *confidence level* of the random forest predictions, i.e., the probability that a sequence belongs to the predicted phenotype (Fig. 2c). LICTOR achieved a specificity of 0.82 and a sensitivity of 0.76 (J=0.58, threshold=0.46 in identifying *tox*), while the *germline-based* classifier showed a 0.69 specificity and a 0.73 sensitivity (J=0.43, threshold=0.48 in identifying *tox*). Overall, these data suggested that SMs harbour key information that can be used to discriminate between *tox* and *nox*, while LC germline VJ rearrangements do not seem to carry additional information that can improve the prediction performance of LICTOR.

LICTOR unveils specific features of LC toxicity

To identify the key features leading to LC toxicity in AL, we ranked the predictor variables of LICTOR according to their "information gain", a value representing the importance of the information carried by each predictor variable for the classification[35]. We found that among the top 10 most important features of the three families of predictor variables, feature 49-A, which denotes an SM to alanine at position 49, obtained the highest score in the AMP family ranking, as well as in the general ranking (Fig. 2d and Supplementary Table 8). Indeed, feature 49-A was present in 54 *tox* sequences but only 8 *nox* sequences. Furthermore, the 49-A mutation, which is located at the dimeric interface of LCs (Fig. 2e), was also ranked among

the top 10 features in the DAP family in combination with no substitutions at other residue positions (Fig. 2d). Moreover, among the best-ranked predictor variables of the three families, those describing mutated positions were more frequent in *tox* sequences than in *nox* sequences (Fig. 2d). Interestingly, all these mutations were located at the LC homodimer interface (Fig. 2e), suggesting that mutations in these positions may affect the structural integrity of the dimeric interface and/or induce local instability of the monomer, thus leading to LC misfolding and aggregation. A similar trend was also observed for other top-ranked features, where unmutated positions were, conversely, more frequent in *nox* sequences than in *tox* sequences (Fig. 2d). Taken together, these findings show that the presence or absence of specific mutations at specific positions of an LC are key features used by LICTOR to classify the LC phenotype. More importantly, this further underlines the pivotal role of SMs as causative of AL.

Reverting the toxic phenotype of an LC using LICTOR

Having assessed the prediction accuracy of LICTOR, we sought to validate the key toxicity determinants identified previously through information gain by computationally reverting the toxic phenotype of an LC with LICTOR and verifying the results in a validated *in vivo C. elegans* model[21,22]. We therefore selected an LC from our database (tox153) previously described in the literature as cardiotoxic[36] and thus having the worst prognosis in AL. For this sequence, we had access to a bone marrow sample, from which we obtained the full-length sequence of tox153 (see Methods). Despite its toxic phenotype, tox153 differs by only 5 SMs from the corresponding germline (Fig. 3b), hence representing a good candidate with which to perform our study. From the analysis of the 5 SMs of tox153, we found that the best candidate, i.e., the mutation with the largest information gain, was at position 52, in which the germline leucine (LEU) was somatically mutated to a valine (VAL). This feature was one of

the top-ranked predictor variables in all three families (AMP, MAP, and DAP). In fact, an unmutated amino acid at position 52 (52X) was significantly more frequent in *nox* sequences than in *tox* sequences (p_{val}=1.5 e-09, Fig. 2d). Therefore, 52X may represent a "non-toxic feature" able to revert the phenotype of tox153 (Fig. 3c). To test this hypothesis, we restored the leucine of the germline sequence at position 52 of tox153 (tox153V52L) and used LICTOR to predict the toxicity of the new sequence. LICTOR predicted tox153V52L as a *nox* sequence, highlighting that this single point mutation is able to completely revert the toxic phenotype of tox153, according to *in silico* prediction. Next, we analysed the other 4 SMs according to their information gain. Among these SMs, only the MAP feature 56X-59X (p_{val} = 2.2 e-16, Fig. 2d) was included in the top ranked. Since tox153 is mutated at position 56 but not at position 59, we also reverted this SM in tox153V52L by mutating alanine to glycine (tox153V52LA56G) (Fig. 3c). Interestingly, *in silico* prediction by LICTOR confirmed the non-toxic phenotype of tox153V52LA56G. These results underline that key predictor variables identified in the feature selection process and used by LICTOR to perform the predictions represent molecular determinants of AL.

Experimental validation of LICTOR using *C. elegans*

Next, we assessed the accuracy of LICTOR's toxicity predictions in a validated *in vivo* model, exploiting the ability of *C. elegans* to specifically identify cardiotoxic LCs[21,22]. To this end, recombinantly expressed tox153, single mutant tox153V52L and double mutant tox153V52LA56G were administered to worms, and their toxicity was evaluated by measuring alterations in the pharyngeal pumping rate[21,22]. Tox153 caused significant pharyngeal dysfunction (Fig. 3d) comparable to that induced by the administration of cardiotoxic LCs purified from patients suffering from AL (H6, H7, and H9) (Fig. 3e) [21,22,37]. The presence of a single mutation in tox153V52L significantly decreased the ability of the

wild-type protein to cause pharyngeal toxicity (Fig. 3d). Notably, the double mutant tox153V52LA56G, similar to LCs purified from patients affected by multiple myeloma (M2, M7, M8)[21,22], did not display toxic activity (Fig. 3e). These results confirm the soundness of LICTOR prediction and the validity of key predictor variables identified by information gain as determinants of *in vivo* proteotoxicity.

Discussion

Early diagnosis of AL is essential to readily apply therapeutic interventions and prevent permanent and fatal organ damage. However, AL is usually detected only once symptoms reflecting advanced organ involvement occur, which results in poor patient prognosis. Moreover, although pre-existing MGUS is a known risk factor for AL, predicting whether MGUS patients will progress to AL remains an open, unsolved problem. The extreme sequence diversity of LCs responsible for AL, due to VJ recombination and SMs, further complicates this scenario. Consequently, to deepen our understanding of the AL determinants and ultimately foster early AL diagnosis, we investigated LC sequences with a known phenotype, with the aim of devising a predictive tool able to flag toxic LCs in advance (i.e., LCs responsible for the formation of toxic aggregates and AL development). To achieve this goal, we analysed a large dataset of toxic (*tox*) and non-toxic (*nox*) LCs of the λ isotype, which is the most frequent isotype in AL, following the hypothesis that specific SMs can increase the propensity of LCs to cause AL, as posed by other research groups[38-40]**.** Therefore, we performed a statistical analysis of the distribution of SMs between *tox* and *nox* sequences. This analysis revealed that toxic LCs have significantly higher SM frequencies than non-toxic LCs (Fig. 1c). Based on these findings, we designed LICTOR, a machine-learning approach using SMs to classify the LC phenotype. LICTOR achieved a specificity and a sensitivity of 0.82 and 0.76, respectively, with an AUC of 0.87, making it an unprecedented tool in early AL

diagnosis. Interestingly, including LC germline VJ rearrangements as additional predictor variables in LICTOR configuration did not improve prediction performance, further suggesting that, despite the prevalence of some VL germline genes in AL, SMs represent the critical driver of the disease.

Next, we showed that specific features (sequence AMP or structural MAP and DAP features) that provide the largest information gain to LICTOR harbour crucial information to accurately predict LC phenotype and can thus be regarded as effective AL molecular determinants. In fact, through the information gain feature selection process, we identified a set of features characterized by the strongest association with the LC phenotype, which, remarkably, were mainly located at the dimeric interface of the LC structure. This finding emphasizes that not only specific SMs but also their structural context are determinants of LC toxicity. The accuracy of LICTOR and of our *in silico* findings was also experimentally confirmed in *C. elegans*, a validated *in vivo* model for measuring LC toxicity. Notably, as predicted by LICTOR, the toxic phenotype of an LC was abolished by reverting a single SM.

Taken together, these findings confirm the accuracy and robustness of our *in silico* approach in the identification of toxic and non-toxic LCs and suggest its usefulness as a diagnostic instrument for AL. Machine learning relies on data. Larger datasets of LC sequences would, therefore, be beneficial for validation, as well as for the improvement of LICTOR accuracy. However, collecting large numbers of toxic LC sequences is difficult due to the low prevalence of the disease. We believe that the application of LICTOR as a diagnostic tool could encourage clinicians to obtain – and make available to the public – LC sequences of AL patients, thus increasing the size of LC sequence databases and consequently allowing improvement of LICTOR accuracy.

In conclusion, LICTOR represents the first method for accurately predicting LC toxicity, allowing the timely identification of high-risk patients, such as MGUS subjects likely to

progress to AL, and thus fostering early treatment and better prognosis. Furthermore, our approach may guide the development of novel predictive tools useful for other diseases, such as cancer, in which the prognosis may depend on SMs of specific tumour-linked proteins. LICTOR is available as a webservice at http:///lictor.irb.usi.ch.

FIGURE 1

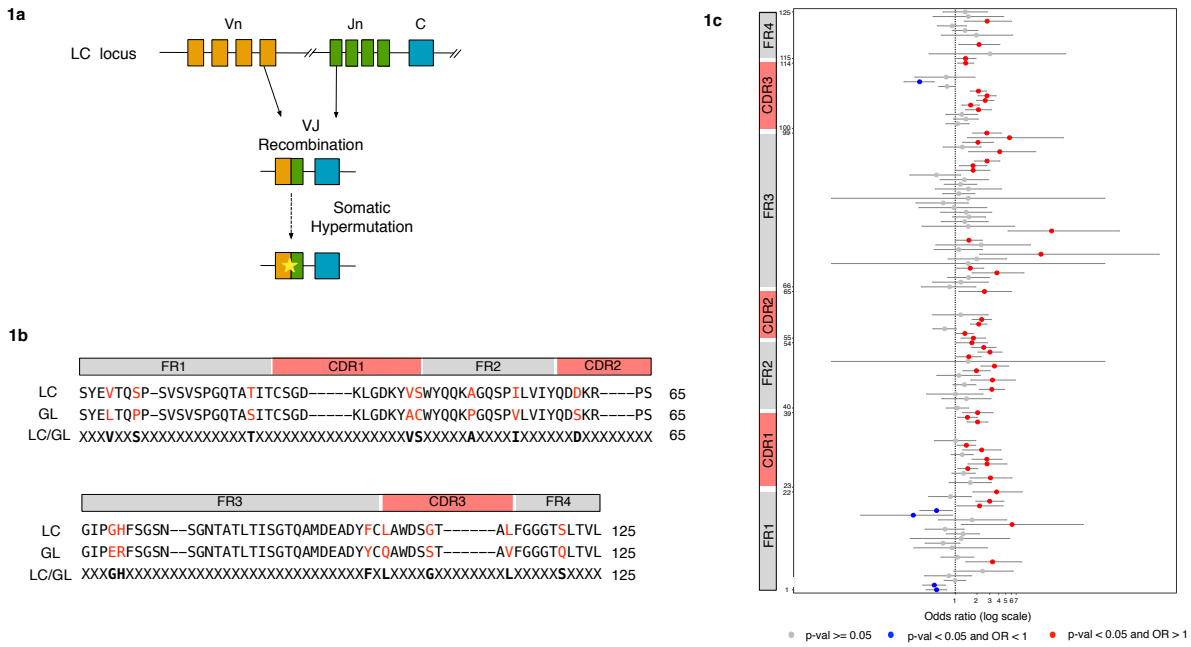
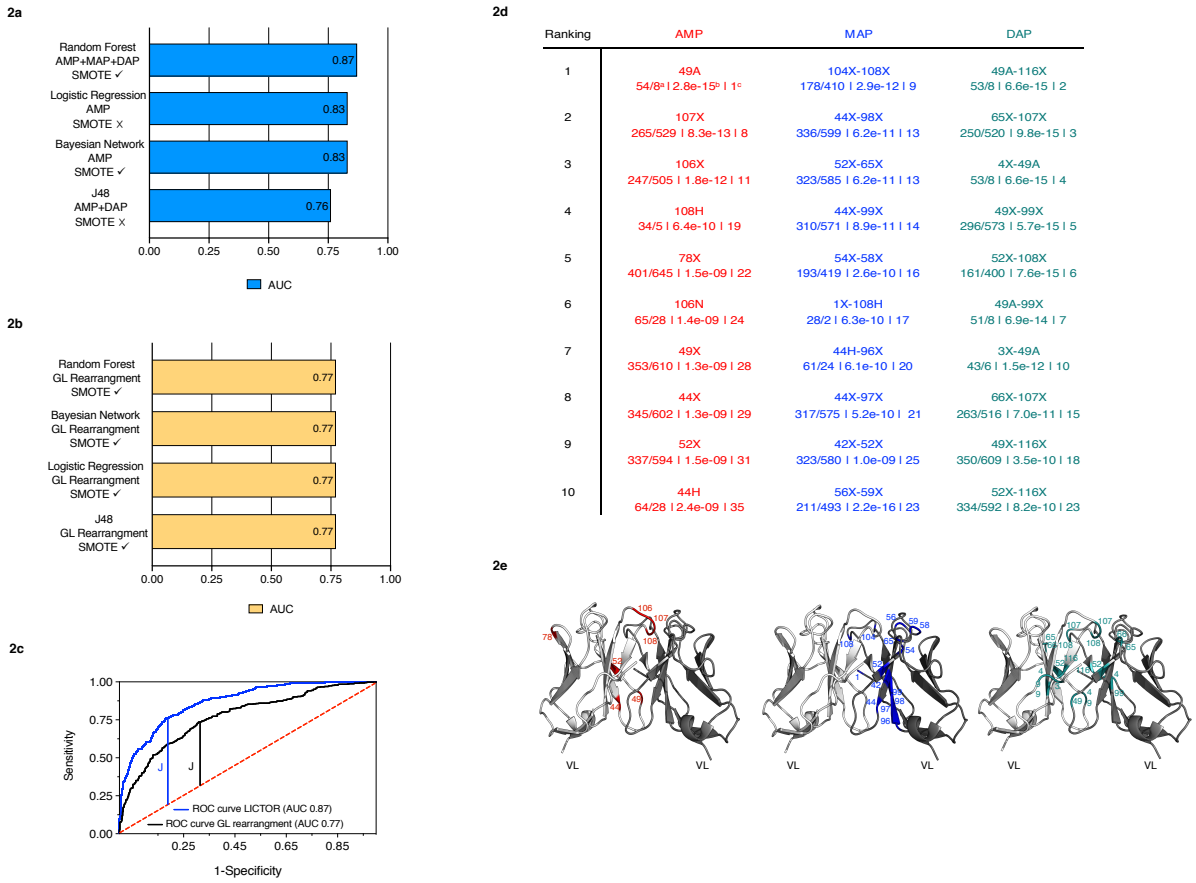


Fig. 1 | The presence of somatic mutations differentiates toxic and non-toxic LC sequences. a, Schematic representation of the generation of LC diversity through the processes of VJ recombination and somatic hypermutation. **b,** Alignment of an LC sequence with the corresponding germline using a progressive Kabat-Chothia numbering scheme with a total of 125 positions. Structural elements of immunoglobulin light chains are depicted on top of the sequences (FR1= framework 1, CDR1= complementary determining region 1, FR2= framework 2, CDR2= complementary determining region 2, FR3= framework 3, CDR3= complementary determining region 3, FR4= framework 4). Residues in red depict somatic mutations. The third line shows the encoding scheme used by the classifier with somatic mutations (displayed in bold) and unmutated positions represented by an 'X'. **c,** OR for all 125 positions of the LC sequences (y-axis). Structural elements of immunoglobulin light chains are shown on the left. ORs for positions with no statistically significant difference between tox and nox sequences ($p \geq 0.05$) are represented as grey dots. Positions with statistically significant differences ($p < 0.05$) are depicted as either red (OR>1) or blue (OR<1) dots. Grey horizontal error bars show the OR 0.95

confidence interval.

FIGURE 2



Ranking	AMP	MAP	DAP
1	49A 54/8[a] I 2.8e-15[b] I 1[c]	104X-108X 178/410 I 2.9e-12 I 9	49A-116X 53/8 I 6.6e-15 I 2
2	107X 265/529 I 8.3e-13 I 8	44X-98X 336/599 I 6.2e-11 I 13	65X-107X 250/520 I 9.8e-15 I 3
3	106X 247/505 I 1.8e-12 I 11	52X-65X 323/585 I 6.2e-11 I 13	4X-49A 53/8 I 6.6e-15 I 4
4	108H 34/5 I 6.4e-10 I 19	44X-99X 310/571 I 8.9e-11 I 14	49X-99X 296/573 I 5.7e-15 I 5
5	78X 401/645 I 1.5e-09 I 22	54X-58X 193/419 I 2.6e-10 I 16	52X-108X 161/400 I 7.6e-15 I 6
6	106N 65/28 I 1.4e-09 I 24	1X-108H 28/2 I 6.3e-10 I 17	49A-99X 51/8 I 6.9e-14 I 7
7	49X 353/610 I 1.3e-09 I 28	44H-96X 61/24 I 6.1e-10 I 20	3X-49A 43/6 I 1.5e-12 I 10
8	44X 345/602 I 1.3e-09 I 29	44X-97X 317/575 I 5.2e-10 I 21	66X-107X 263/516 I 7.0e-11 I 15
9	52X 337/594 I 1.5e-09 I 31	42X-52X 323/580 I 1.0e-09 I 25	49X-116X 350/609 I 3.5e-10 I 18
10	44H 64/28 I 2.4e-09 I 35	56X-59X 211/493 I 2.2e-16 I 23	52X-116X 334/592 I 8.2e-10 I 23

Fig. 2 | Machine learning predicts toxic and non-toxic sequences and identifies key features of toxicity. a, AUC of the best configuration for each of the considered machine learners (blue bars). Different combinations of three families of predictor variables were tested, with (✓) or without (✗) the SMOTE balancing technique. **b,** The yellow bars show the best AUC value obtained by each machine learner using only the LC germline VJ rearrangements as predictor variables. **c,** ROC curve for LICTOR (i.e., random forest using AMP + MAP + DAP) compared with a predictor (random forest) using only the LC germline VJ rearrangements as predictor variables. **d,** Top 10 features of each family ranked by information gain. Below each predictor variable are shown the occurrence in tox/nox sequences ([a]), the p-value ([b]) and the feature selection general ranking ([c]) (red= AMP features, blue= MAP features, and green=DAP features). **e,** Mapping of the top 10 features of each family on the variable domains of an LC homodimeric structure (PDB ID: 2OLD, represented in white and grey in cartoon). AMP features are shown in red in the left image, MAP features in blue in the middle image,

and DAP features in green in the right image. The colour code used in the table to represent the three feature families is maintained in their structural representation in d.

FIGURE 3

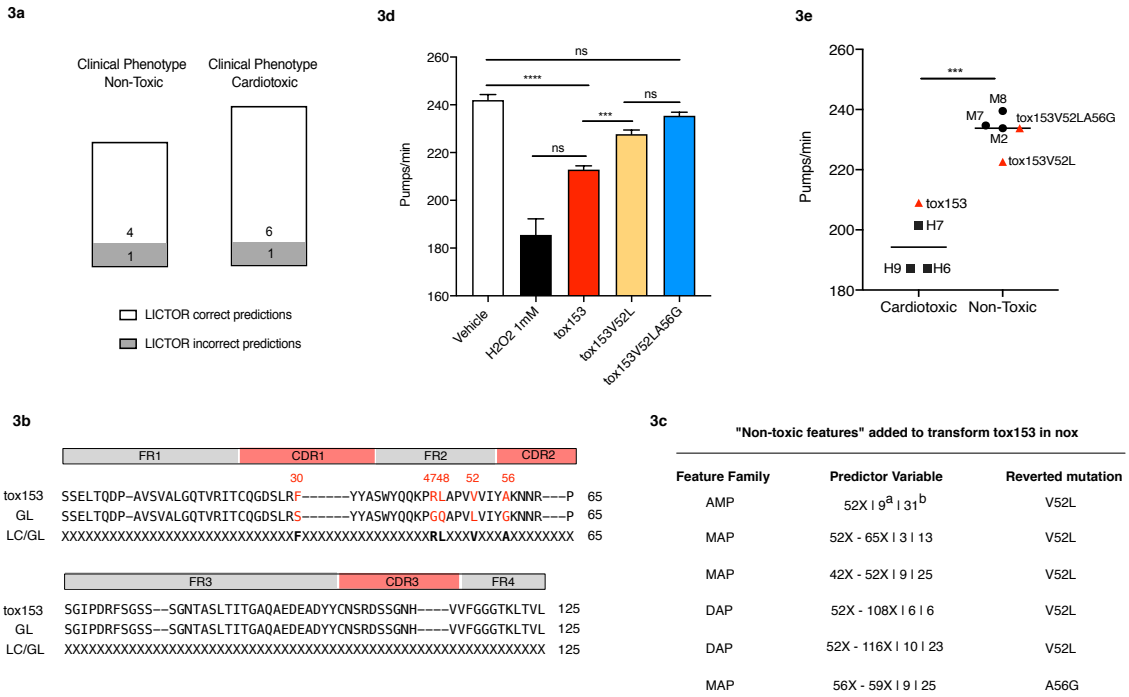


Fig. 3 | LICTOR accurately predicts the LC toxicity of sequences absent from the training set and is able to revert the pathological phenotype of a cardiotoxic LC. a, LICTOR predictions based on an independent set of LCs, i.e., not present in the training set. Toxic LCs are from patients affected by AL with cardiac involvement, while non-toxic LCs are from patients with multiple myeloma (see also Supplementary Table 2). Predictions are divided according to the clinical phenotype. White bars represent correct LICTOR predictions, while grey bars represent incorrect predictions. **b,** Sequence of a cardiotoxic LC (tox153) used to neutralize the toxic phenotype using LICTOR and the non-toxic features unveiled by feature selection. Tox153 is aligned with the corresponding germline (GL) sequence, and the third line shows the difference in somatic mutations between the LC and the GL sequence (LC/GL) with somatic mutations (displayed in bold) and unmutated positions represented by an 'X'. **c,** The table represents the non-toxic features according to information gain used to revert the toxic phenotype of tox153. For each predictor variable, we also report the ranking in the specific feature

family ([a]) and the feature selection general ranking ([b]) **d,** Proteotoxic effect of tox153 protein and of the two mutants *in silico* designed by adding non-toxic features (tox153V52L and tox153V52LA56G). Proteins in 10 mM PBS (100 μg/ml) were administered to *C. elegans* (100 μl/100 worms). Vehicle (10 mM PBS) and 1 mM H_2O_2 were administered as negative and positive controls, respectively. Pharyngeal activity was determined 24 h after treatment by determining the number of pharyngeal bulb contractions (pumps/min). Data are the mean pumps/min ± SE (n= 30 worms/assay, two assays). **** $p < 0.0001$, *** $p < 0.001$ one-way ANOVA, Dunn's *post hoc* test. **e,** Values of pharyngeal bulb contraction (pumps/min) of some LCs listed in Supplementary Table 2 are from AL patients with cardiac involvement (Cardiotoxic) and patients with multiple myeloma (Non-Toxic). These values were previously obtained under the same experimental conditions employed in this study[21,22]. Each square is the mean value for H6, H7 and H9, while dots represent the mean of M2, M7 and M8. Horizontal lines represent the mean of cardiotoxic and non-toxic LCs (*** $p < 0.0004$, unpaired t-test). The values of pumps/min obtained after the administration of tox153, tox153V52L, and tox153V52LA56G are also plotted (red triangles).

Methods

Dataset. The database used in the training was composed of 428 *tox* and 590 *nox* sequences of the λ isotype collected from the Amyloid Light-chain Database (ALBase) (http://albase.bumc.bu.edu). Furthermore, it contained 57 *nox* λ LC sequences that we collected at the Institute for Research in Biomedicine (IRB-DB), known to be non-toxic in the context of AL. The 1,075 sequences were automatically aligned using a progressive Kabat-Chothia numbering scheme (http://www.bioinf.org.uk/abs/). According to this scheme, for example, the CDR1 of a given LC with Kabat-Chothia numbering 30A, 30B, 30C, 30D, 30E, and 30F was assigned 31, 32, 33, 34, 35, and 36. For the ALBase sequences, germline information was taken from the database, while for IRB-DB LCs, the germline was assessed

with an in-house script. Next, germline (GL) sequences were reconstructed using the IMGT database[25].

The GL sequences were aligned with the same numbering scheme used for the LCs. Next, each LC in the dataset was compared with the corresponding GL to identify all somatic mutations, with the differences encoded using an X for unmutated positions and the LC amino acid for somatic mutations; this sequence was referred to as S_{mut}. For example, an LC with sequence *SYELTQPP* and a corresponding GL with the sequence *SYVLTQPP* was encoded as *XXEXXXXX* since there is a somatic mutation (*V*→*E*) at position 3. To compare the presence of somatic mutations in S_{mut} at each position *i* in the Kabat-Chothia numbering scheme, the following four quantities were computed:

- tox^i_{NM} - the number of toxic sequences without a somatic mutation at position *i;*
- tox^i_M - the number of toxic sequences with a somatic mutation at position *i;*
- nox^i_{NM} - the number of non-toxic sequences without a somatic mutation at position *i;*
- nox^i_M - the number of non-toxic sequences with a somatic mutation at position *i.*

Statistical analysis. The *fisher.test* function in R version 3.5.1 with the arguments *conf.int=TRUE* and *conf.level*=0.95 was used to assess significant differences in somatic mutations between toxic and non-toxic sequences. The OR between tox^i_M/tox^i_{NM} and nox^i_M/nox^i_{NM} was computed as

$$OR^i_{tox\text{-}nox} = \frac{tox^i_M / tox^i_{NM}}{nox^i_M / nox^i_{NM}}$$

OR=1 indicates that the event under study (i.e., the frequency of mutations at position *i*) is equally likely in the two groups (e.g., *tox* vs *nox*). OR>1 indicates that the event is more likely in the first group (*tox*). OR<1 indicates that the event is more likely in the second group (*nox*). The *t.test* function in R version 3.5.1 was used to evaluate whether the probability distributions of the number of SMs (PDSM) differed between the *tox*, *nox* and *hdnox* datasets. Data from *C. elegans*-based assays were analysed using GraphPad Prism 8.2.1 software by one-way ANOVA and Dunn's post-test analysis. A P value < 0.05 was considered significant.

Predictor variables used by the machine learners. Given a sequence, the following features were extracted:

Amino acid at each Mutated Position (AMP). From a sequence S_{mut}, a list of predictor variables was extracted, each describing the type of amino acid added by the somatic mutation at a given position or the absence of a mutation at the position. Thus, each of these variables was a pair (*position*, *amino acid*), where we used the letter "X" instead of the amino acid at the positions for which no somatic mutations were present.

Monomeric Amino acid Pairs (MAP). LCs share a conserved 3D structure. Therefore, pairs of interacting residues were defined as amino acids having a distance between the respective Cβ atoms less than 7.5 Å in the X-ray structure (PDB ID: 2OLD).

Dimeric Amino acid Pairs (DAP). Similarly, pairs of residues that interact at the LC-LC interface were defined using the 2OLD LC homodimeric X-ray structure. Two residues belonging to different chains were considered to interact if the distance between their Cβ atoms was less than 7.5 Å.

Machine learning algorithms. Weka 3.8.1[30] implementation was used for the four machine learning algorithms (Bayesian network, logistic regression, J48, and random forest) to solve

the classification task. For all algorithms, the default Weka parameters were used. The algorithms were evaluated by performing 10-fold cross-validation over the dataset. The performance of each algorithm was first assessed using only one family of features (e.g., AMP, MAP, and DAP, for a total of three combinations); second, the three families were combined into pairs (e.g., AMP U MAP, for a total of three combinations); third, all three families were combined together. This led to a total of 7 (feature configurations) × 4 (algorithms) = 28 prediction experiments. Moreover, each of the 28 experiments was performed with and without the balancing of the training set with SMOTE (**S**ynthetic **M**inority **O**ver-sampling **TE**chnique)[31] on the toxic sequences so that the number of toxic instances was equal to the number of non-toxic instances in the training set during each of the ten cross-validations used in the evaluation. This led to 28 × 2 (with/without SMOTE) = 56 total experiments.

Prediction performance. The various prediction algorithms were assessed by computing the following classification errors: (i) Type I misclassifications, indicating toxic sequences incorrectly classified as non-toxic (false negative—FN), and (ii) Type-II misclassifications, indicating non-toxic sequences misclassified as toxic (false positive—FP). The correct classifications were instead indicated by the number of true positives—TP (a toxic sequence correctly classified) —and true negatives—TN (a non-toxic sequence correctly classified). Based on TP, TN, FP, and FN, the following metrics were used to evaluate the performance of our classifiers:

- *Area under the receiver operating characteristic curve (AUC).* The AUC is used to assess the performance of a two-class classifier (such as that in our study) and is equal to the probability that the classifier will rank a randomly chosen positive instance (in our case, a toxic sequence) higher than a randomly chosen negative instance (a non-toxic sequence). A random classifier has an AUC=0.5, while the AUC is 1.0 for a perfect classifier.

- *Sensitivity.* Computed as TP/(TP+FN), this represents the percentage of toxic sequences correctly identified by the classifier.
- *Specificity.* Computed as TN/(TN+FP), this represents the percentage of non-toxic sequences correctly identified by the classifier.

Youden index.

The Youden (J) index was used to validate the effectiveness of the predictors and to find the optimal cut-off point to separate toxic LCs associated with the disease from non-toxic LCs using the following formula:

$J=\max_c[Se(c) + Sp(c) -1]$

Information gain feature selection. The InfoGainAttributeEval filter implemented in Weka 3.8.1[35] was used to remove all features that did not contribute to the information available for the prediction of the sequence type. All features having an information gain less than 0.01 were removed. Given the computational cost of this procedure, this experiment was performed for the best-performing algorithm and configuration identified in the previous 56 experiments. The full list of ranked features is shown in Supplementary Table 8.

Tox153 sequence

The full-length LC sequence was obtained from bone marrow cells with TRIzol lysate. RNA was extracted using Direct-zol RNA Miniprep Zymo-Spin IIC columns (Zymo Research) and used for cDNA synthesis, employing the template switching technique approach[41]. Ig genes were amplified with specific primers, and a library of unique sequences was obtained by a Zero Blunt® TOPO® PCR Cloning Kit (Life Technologies) and subsequent analysis of single colonies.

Protein production and purification

Tox153, tox153V52L and tox153V52LA56G were custom expressed in mammalian cell lines (Expi293F), purified by CaptureSelect™ LC-lambda (Hu) Affinity Matrix, and analysed by SDS-Page and Western blot by GenScript (New Jersey, USA).

Effect of LCs on *C. elegans*

Bristol N2 nematodes were obtained from the *Caenorhabditis elegans* Genetic Center (CGC, University of Minnesota, Minneapolis, MN) and propagated at 20°C on solid nematode growth medium (NGM) seeded with *Escherichia coli* OP50 (CGC) for food. Worms were incubated with 100 µg/ml tox153 wild-type protein, tox153V52L or tox153V52LA56G (100 worms/100 µl) in 10 mM phosphate-buffered saline (PBS, pH 7.4)[8,9]. Hydrogen peroxide (1 mM) was administered under dark conditions as a positive control and 10 mM PBS (pH 7.4) as a negative control (vehicle). After 2 h of incubation with orbital shaking, worms were transferred onto NGM plates seeded with OP50 *E. coli*. The pharyngeal pumping rate, measured by counting the number of times the terminal bulb of the pharynx contracted over a 1-min interval, was scored 24 h later.

Acknowledgements

This study was supported by grants from the Swiss National Science Foundation (31003A-166472) to A.C. and from the Italian Ministry of Health (RF-2013-02355259 and RF-2016-02361756) to L.D. *C. elegans* and OP50 *E. coli* were provided by the CGC, which is funded by NIH Office Research Infrastructure Programs (P40 OD010440). We would like to acknowledge the use of the Boston University ALBase, supported by HL68705, and Dr. Giampaolo Merlini and Dr. Mario Nuvolone, who provided bone marrow samples.

Author contributions

A.C. and M.G. designed the research. M.G. performed data acquisition and analysis and drafted the manuscript. M.F., L.P., S.R., M.R. and A.L. provided data. M.G., L.P., S.R., M.F., M.M., J.S., M.P., O.M., A.L., L.D. and A.C. edited and approved the manuscript.

References

1 Merlini, G. & Bellotti, V. Molecular mechanisms of amyloidosis. *N Engl J Med* **349**, 583-596, doi:10.1056/NEJMra023144 (2003).
2 Lavatelli, F. *et al.* A novel approach for the purification and proteomic analysis of pathogenic immunoglobulin free light chains from serum. *Biochim Biophys Acta* **1814**, 409-419, doi:10.1016/j.bbapap.2010.12.012 (2011).
3 Lousada, I., Comenzo, R. L., Landau, H., Guthrie, S. & Merlini, G. Light Chain Amyloidosis: Patient Experience Survey from the Amyloidosis Research Consortium. *Adv Ther* **32**, 920-928, doi:10.1007/s12325-015-0250-0 (2015).
4 Palladini, G. & Merlini, G. What is new in diagnosis and management of light chain amyloidosis? *Blood* **128**, 159-168, doi:10.1182/blood-2016-01-629790 (2016).
5 Merlini, G. *et al.* Systemic immunoglobulin light chain amyloidosis. *Nat Rev Dis Primers* **4**, 38, doi:10.1038/s41572-018-0034-3 (2018).
6 Grogan, M., Dispenzieri, A. & Gertz, M. A. Light-chain cardiac amyloidosis: strategies to promote early diagnosis and cardiac response. *Heart* **103**, 1065-1072, doi:10.1136/heartjnl-2016-310704 (2017).
7 Gertz, M. A. Immunoglobulin light chain amyloidosis diagnosis and treatment algorithm 2018. *Blood Cancer J* **8**, 44, doi:10.1038/s41408-018-0080-9 (2018).
8 Merlini, G. Determining the significance of MGUS. *Blood* **123**, 305-307, doi:10.1182/blood-2013-12-539940 (2014).
9 Blancas-Mejia, L. M. & Ramirez-Alvarado, M. Systemic amyloidoses. *Annu Rev Biochem* **82**, 745-774, doi:10.1146/annurev-biochem-072611-130030 (2013).
10 Wang, P. *et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* **68**, 1813-1819, doi:10.1136/gutjnl-2018-317500 (2019).
11 Shen, L. *et al.* Deep Learning to Improve Breast Cancer Detection on Screening Mammography. *Sci Rep* **9**, 12495, doi:10.1038/s41598-019-48995-4 (2019).
12 Yoo, S., Gujrathi, I., Haider, M. A. & Khalvati, F. Prostate Cancer Detection using Deep Convolutional Neural Networks. *Sci Rep* **9**, 19518, doi:10.1038/s41598-019-55972-4 (2019).
13 Saltz, J. *et al.* Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep* **23**, 181-193 e187, doi:10.1016/j.celrep.2018.03.086 (2018).
14 Corredor, G. *et al.* Spatial Architecture and Arrangement of Tumor-Infiltrating Lymphocytes for Predicting Likelihood of Recurrence in Early-Stage Non-Small Cell Lung Cancer. *Clin Cancer Res* **25**, 1526-1534, doi:10.1158/1078-0432.CCR-18-2013 (2019).
15 Olivecrona, M., Blaschke, T., Engkvist, O. & Chen, H. Molecular de-novo design through deep reinforcement learning. *J Cheminform* **9**, 48, doi:10.1186/s13321-017-0235-x (2017).
16 Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol Pharm* **14**, 3098-3104, doi:10.1021/acs.molpharmaceut.7b00346 (2017).
17 Costello, J. C. *et al.* A community effort to assess and improve drug sensitivity prediction algorithms. *Nat Biotechnol* **32**, 1202-1212, doi:10.1038/nbt.2877 (2014).
18 Ali, M., Khan, S. A., Wennerberg, K. & Aittokallio, T. Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach. *Bioinformatics* **34**, 1353-1362, doi:10.1093/bioinformatics/btx766 (2018).

19 Cortes-Ciriano, I. *et al.* Improved large-scale prediction of growth inhibition patterns using the NCI60 cancer cell line panel. *Bioinformatics* **32**, 85-95, doi:10.1093/bioinformatics/btv529 (2016).
20 Rajkomar, A., Dean, J. & Kohane, I. Machine Learning in Medicine. *N Engl J Med* **380**, 1347-1358, doi:10.1056/NEJMra1814259 (2019).
21 Diomede, L. *et al.* A Caenorhabditis elegans-based assay recognizes immunoglobulin light chains causing heart amyloidosis. *Blood* **123**, 3543-3552, doi:10.1182/blood-2013-10-525634 (2014).
22 Diomede, L. *et al.* Cardiac Light Chain Amyloidosis: The Role of Metal Ions in Oxidative Stress and Mitochondrial Damage. *Antioxid Redox Signal* **27**, 567-582, doi:10.1089/ars.2016.6848 (2017).
23 Bodi, K. *et al.* AL-Base: a visual platform analysis tool for the study of amyloidogenic immunoglobulin light chain sequences. *Amyloid* **16**, 1-8, doi:10.1080/13506120802676781 (2009).
24 Dispenzieri, A., Gertz, M. A. & Buadi, F. What do I need to know about immunoglobulin light chain (AL) amyloidosis? *Blood Rev* **26**, 137-154, doi:10.1016/j.blre.2012.03.001 (2012).
25 Brochet, X., Lefranc, M. P. & Giudicelli, V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Res* **36**, W503-508, doi:10.1093/nar/gkn316 (2008).
26 Sheskin, D. *Handbook of parametric and nonparametric statistical procedures*. 4th edn, (Chapman & Hall/CRC, 2007).
27 DeKosky, B. J. *et al.* Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proc Natl Acad Sci U S A* **113**, E2636-2645, doi:10.1073/pnas.1525510113 (2016).
28 Makino, D. L., Henschen-Edman, A. H., Larson, S. B. & McPherson, A. Bence Jones KWR protein structures determined by X-ray crystallography. *Acta Crystallogr D Biol Crystallogr* **63**, 780-792, doi:10.1107/S0907444907021981 (2007).
29 Oberti, L. *et al.* Concurrent structural and biophysical traits link with immunoglobulin light chains amyloid propensity. *Sci Rep* **7**, 16809, doi:10.1038/s41598-017-16953-7 (2017).
30 Hall M., F. E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.* **11**, 10-18 (2009).
31 Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res* **16**, 321-357, doi:DOI 10.1613/jair.953 (2002).
32 Abraham, R. S. *et al.* Immunoglobulin light chain variable (V) region genes influence clinical presentation and outcome in light chain-associated amyloidosis (AL). *Blood* **101**, 3801-3808, doi:10.1182/blood-2002-09-2707 (2003).
33 Comenzo, R. L., Zhang, Y., Martinez, C., Osman, K. & Herrera, G. A. The tropism of organ involvement in primary systemic amyloidosis: contributions of Ig V(L) germ line gene use and clonal plasma cell burden. *Blood* **98**, 714-720, doi:10.1182/blood.v98.3.714 (2001).
34 Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32-35, doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3 (1950).
35 Witten, I. H., Frank, E. & Hall, M. A. *Data mining : practical machine learning tools and techniques*. 3rd edn, (Morgan Kaufmann, 2011).

36 Perfetti, V. *et al.* The repertoire of lambda light chains causing predominant amyloid heart involvement and identification of a preferentially involved germline gene, IGLV1-44. *Blood* **119**, 144-150, doi:10.1182/blood-2011-05-355784 (2012).
37 Maritan, M. *et al.* Inherent Biophysical Properties Modulate the Toxicity of Soluble Amyloidogenic Light Chains. *J Mol Biol*, doi:10.1016/j.jmb.2019.12.015 (2019).
38 Baden, E. M., Randles, E. G., Aboagye, A. K., Thompson, J. R. & Ramirez-Alvarado, M. Structural insights into the role of mutations in amyloidogenesis. *J Biol Chem* **283**, 30950-30956, doi:10.1074/jbc.M804822200 (2008).
39 del Pozo Yauner, L. *et al.* Influence of the germline sequence on the thermodynamic stability and fibrillogenicity of human lambda 6 light chains. *Proteins* **72**, 684-692, doi:10.1002/prot.21934 (2008).
40 Davis, D. P. *et al.* Both the environment and somatic mutations govern the aggregation pathway of pathogenic immunoglobulin light chain. *J Mol Biol* **313**, 1021-1034, doi:10.1006/jmbi.2001.5092 (2001).
41 Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-181, doi:10.1038/nprot.2014.006 (2014).