Data-driven Discovery of Potent Small Molecule Ice Recrystallisation Inhibitors

Matthew T. Warren[1,2], Caroline I. Biggs[1], Akalabya Bissoyi[2], Matthew I. Gibson[1,2*] and Gabriele C. Sosso[1*]

[1]Department of Chemistry, University of Warwick, Coventry, CV4 7AL, United Kingdom.
[2]Warwick Medical School, University of Warwick, Coventry, CV4 7AL, United Kingdom.

*Corresponding author(s). E-mail(s): m.i.gibson@warwick.ac.uk; g.sosso@warwick.ac.uk;

Abstract

Controlling the formation and growth of ice is essential to successfully cryopreserve cells, tissues and biologics for research or clinical use. Current programmes to identify materials capable of modulating ice growth are guided solely by iterative changes and human intuition, with a major focus on macromolecules (proteins or polymers). This process is fundamentally constrained by a poor understanding of the mechanisms and underlying structure-activity relationships. Here, we overcome this barrier by constructing machine learning models capable of predicting the ice growth inhibition activity of small molecules. Due to current limitations in experimental throughput, we leverage ensemble models which combine state-of-the-art descriptors with domain-specific features derived from molecular simulations. When applied to virtually screen a commercial compound library, these models successfully predicted novel ice recrystallisation inhibitors that are experimentally verified to function at low millimolar concentrations. This data-driven approach will enable the discovery of new cryoprotectants to address the rapidly growing clinical and biotechnological cold-chain demands.

Keywords: ice, ice recrystallisation inhibitors, machine learning, data-driven discovery, cryopreservation

1 Introduction

The low-temperature storage of cells, tissues and biologics is an essential technology for biomedical research, regenerative medicine, and vaccination distribution [1, 2]. These materials are most effectively preserved at subzero temperatures, however there are several stressors associated with these conditions, originating from the formation (nucleation) and growth (recrystallisation) of ice [3]. In nature, many organisms can mitigate the harmful effects of ice by accumulating solutes which provide a cryoprotective effect on a colligative basis [4]. Cold-adapted species may also produce antifreeze proteins and macromolecules which can recognise and bind to nascent ice crystals directly, inhibiting growth at the bound surface [5]. This process, known as ice recrystallisation inhibition (IRI), has drawn significant interest due to its potential to enhance the cryopreservation of cells and tissues *ex vivo*. To this end, a variety of biomimetic materials have been explored, including synthetic protein analogues [6], polymers [7] and self-assembling compounds [8]. More recently, small molecules have also been found to slow the recrystallisation of ice via an alternative mechanism that does not involve binding directly to the crystal surface, improving the cryopreservation outcomes for human blood and stems cells, as well as mammalian organs [9–11]. Nonetheless, these protocols still required significant quantities of cosolvent (e.g. DMSO, glycerol) which are associated with several adverse effects, including toxicity and epigenetic alterations [12, 13]. Moreover, the diversity of cell characteristics and storage requirements means that rarely does any single cryoprotectant provide optimal outcomes across different settings [3, 14]. Consequently, there is a pressing need for novel ice recrystallisation inhibitors (IRIs) for a number of applications, most notably cell-based therapies which require cryopreservation throughout their supply chains. [2].

Despite significant efforts, identifying new IRIs remains a formidable task. Guided by limited mechanistic understanding, as well as structure-activity relationships that can be counterintuitive and contradictory, the discovery of new materials is fundamentally trial and error. Traditionally, the synthesis and screening of tens to hundreds of compounds has been required in order to identify a handful of potent "hits" [15]. This process is time-intensive, compounded by a lack of high-throughput techniques to quantitatively assess IRI, and has thus become the bottleneck in the development of IRI-active molecules as next-generation cryoprotectants. In our previous work, we introduced amino acids as a novel class of IRIs that can be highly effective at millimolar concentrations [16, 17]. Benefiting from their low-cost, chemical diversity and commercial availability, amino acids are ideally suited for wider study, and poised for industrial application as cryoprotectants. In this article, we introduce a machine learning-based approach for predicting the IRI activity of amino acids. Our method combines a number of state-of-the-art descriptors, as well as novel representations computed via atomistic molecular dynamics (MD) simulations, targeting interactions between hydrated inhibitors and

water. To overcome the challenges of low-throughput experimental measurements of IRI, we leverage this descriptive information by using an ensemble of models with different inputs and architectures, to obtain more robust predictions and quantify prediction uncertainty. We show for the first time that these models can be used to accurately predict IRI activity and apply this to screen a commercial compound library. These predictions were experimentally validated, discovering non-obvious small molecule IRIs. Overall, this data-driven approach signals a fundamental shift towards the rapid and efficient discovery of IRI-active materials for applications in industrial and clinical sciences. We discuss these findings and evaluate the significance of molecular hydration as a determinant of small molecule IRI activity.

2 Results and discussion

2.1 Classifying IRI-active small molecules

To benchmark our methods against earlier work, we first developed classification models to predict IRI activity on a categorical basis. Previously, Briard and coworkers [15] constructed a binary classification model to predict whether small-molecule carbohydrates were IRI-active or IRI-inactive using the training set denoted here as the *Glyco* dataset (Table 1, Figure 1a). IRI activity was measured using the "splat cooling" assay, wherein the average ice crystal (grain) size is determined from a micrograph following a period of recrystallisation (Figure 1d), as described in the Methods. This assay yields data in the form of mean grain size (MGS) values, which are typically normalised against a negative control for inhibition, producing a relative (%) MGS metric. A smaller % MGS value thus signifies stronger inhibition (Figure 1d).

Table 1 Datasets featured in this work. Complete datasets including compound names, SMILES and % MGS values can be found in the GitHub repository [18]

Dataset	No. compounds	Compound classes
Glyco[a]	124	Aryl glycosides, aryl/alkyl aldonamides
Glyco2[b]	223	Aryl/allyl/amino glycosides, aryl/alkyl aldonamides mono/disaccharides, lysine-based surfactants cationic anti-INAs
Amino	63	(α/β/γ-)amino acids/alcohols/esters
Combined[c]	286	All of the above

[a]Same dataset used in Ref. [15]. [b]Includes compounds in Glyco. [c]Includes compounds in Glyco2 and Amino.

For classification, % MGS values must therefore be converted into categorical ("active" and "inactive") labels according to given threshold, chosen to be 70 % MGS by Briard and colleagues [15]. To construct classification models, we computed six different molecular descriptors for each structure, encompassing

low- and high-dimensional representations, as detailed in the Methods. Neural network (NN) classifiers were then trained, optimised and validated for the *Glyco* dataset using the architectures and training/validation procedures given in the Methods and Supplementary Information (Supplementary Table 3). We note briefly here that, given the small size of the datasets, a leave-one-out (LOO) cross validation (CV) procedure has been applied to obtain robust predictions for each molecule which are independent of training and test splits. Model performance was evaluated using the same metrics as Briard et al. [15], defined in the Supplementary Information. The performance of the classification models in terms of these metrics are displayed in Table 2. These metrics are computed on test predictions obtained via a single LOO CV run, however we found that repeating this process generates very similar scores (Supplementary Table 7), hence we report only the values for a single representative model. Overall, these models performed reasonably well, with most achieving an F-score of 0.64 or above, similar to the benchmark score of 0.67 previously set. The only exception to this was the classifier trained using the hydration indices, which had an F-score of 0.50 and poor sensitivity of 0.38. Conversely, the best performing model based on the F-score was the molecular cliques, however this model also had low specificity.

Given that we had a number of individual models, each showing satisfactory performance having been trained using a single descriptor, we sought to improve the predictions by aggregating the model outputs to obtain a consensus, as illustrated in Figure 1e. This approach, also known as ensemble learning, is based on a simple rationale: combining multiple models each with intrinsic variability and weakness can result in more robust and accurate predictions. This is fundamental to certain supervised learning algorithms (e.g. random forests [19]) and has also proved successful for chemical property prediction [20]. With this in mind, ensemble classification predictions were obtained using a majority-voting scheme, whereby the most frequently predicted class among a set of models is taken as the consensus. We explored all sets of three or more different descriptor models, ranking them based on their F-scores. The results of the top three performing ensembles, shown in Table 2, revealed improvements across most metrics compared to any individual model. Across these metrics, the best results were obtained predictions from all six descriptors were combined (Ensemble 2, Table 2), despite the fact that contributing models (e.g. those built using H-wACSFs, see Methods), had weak performance individually. Although the prediction specificity (0.68) was lower compared to other individual models, the model's precision (0.82) represents an improvement compared to the previous benchmark. We also emphasise that high precision can be advantageous in applications such as this, wherein the primary objective is to accelerate the discovery of new IRI-active materials.

Despite these encouraging results, we argue that predicting whether a given molecule belongs to the "active" or "inactive" class is not the optimal framework for this task. Typically, classification models are used when the target property is categorical by nature. In this case, as the target is a numerical

quantity, an activity threshold must be defined *a priori*. As there is no clear definition for an "active" small molecule IRI, the modeller's choice of threshold appears arbitrary, yet in fact we show this is crucial, having a significant impact on the results by adjusting the proportion of class labels in training data. In Figure 1b and c, we highlight the effects of different thresholds on the numbers of "active" and "inactive" molecules, and the performance of the corresponding ensemble model featuring all descriptors. Across these thresholds, the best scores were obtained using a value of 70 % MGS, which was also the threshold used in the previous benchmark [15], resulting in a slight imbalance where there are more active than inactive compounds (Figure 1b) in the dataset. Although this model had good predictive power, it has limited practical utility, providing no way of distinguishing highly active structures (i.e. % MGS < 20) from moderately active ones (% MGS > 40), bearing in mind that only highly active are of interest with respect to use as cryoprotectants. Instead, we suggest that this task is better formulated as a regression problem, wherein the model is trained to predict absolute % MGS values.

Table 2 Performance metrics for test predictions obtained from classification models using an activity threshold of 70 % MGS.

Model	Sensitivity	Specificity	Precision	F-score
Benchmark[a]	0.67	0.80	0.67	0.67
Standard descriptors	0.59	0.66	0.76	0.67
Molecular cliques	0.78	0.46	0.72	0.74
H-wACSFs	0.62	0.43	0.66	0.64
SOAPs	0.65	0.55	0.72	0.68
Hydration histograms	0.54	0.71	0.77	0.64
Hydration indices	0.38	0.77	0.75	0.50
Ensemble 1[b]	0.68	0.68	0.79	0.74
Ensemble 2[c]	0.62	0.75	0.82	0.71
Ensemble 3[d]	0.67	0.61	0.76	0.71

[a]Data obtained from Ref.[15]. [b]Ensemble using all descriptors except hydration histograms. [c]Ensemble using all descriptors. [d]Ensemble using all descriptors except hydration histograms and indices.

2.2 Predicting absolute IRI activity

Stepping away from the classification framework described in the previous section, we trained and evaluated NN regressors using each descriptor independently, as described in the Methods. At this point, we switched to using the *Glyco2* dataset, which includes the entire *Glyco* set, as well as 99 additional carbohydrates tested for IRI activity under the same conditions (Table 1). These structures are sufficiently similar to enable efficient training, while also increasing the chemical diversity of the dataset (Supplementary Figure 1). In

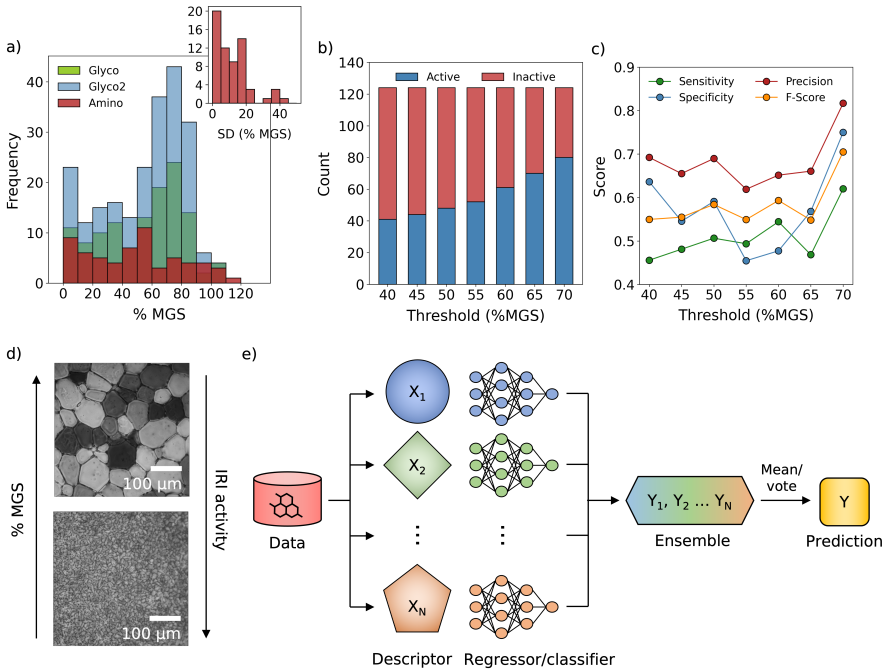

Fig. 1 a) Frequency histograms showing the distribution of % MGS values across the different datasets. Inset shows the distribution of error (SD) for the *Amino* dataset. b) The effects of the activity threshold on the proportion of compounds labelled "active" or "inactive" in the *Glyco* dataset. c) Performance of the ensemble classification model trained and tested on the *Glyco* dataset, comprising all (six) descriptors. d) Cryomicrographs obtained from the "splat cooling" assay performed on an IRI-inactive (top) and IRI-active (bottom) material. e) Schematic illustrating the ensemble learning approach. Different descriptors are computed and used to train independent models, whose outputs are combined via majority voting scheme (classification) or mean average (regression).

addition, we have also assembled a new dataset – *Amino* – via our own experimental measurements of 63 amino acids, shown previously to be a novel class of IRI-active material [16, 17] (Table 1). IRI activity was evaluated using the "splat cooling" assay, however different solution conditions were used which can affect the magnitude of observed activity. This issue is discussed in more detail in the Methods.

Separate regressors were trained and evaluated using the *Amino* and *Glyco2* datasets independently and in combination (see *Combined* dataset). The results of these models, reported using mean squared error (MSE) and Pearson's correlation coefficient (PCC) metrics, are shown in Table 3. As with the classification models, we report the metrics computed on a complete set of test predictions generated via LOO CV, noting that very similar results are produced if this process is repeated (Supplementary Table 6). Individually, most models had a relatively high predictive error, with experimental and predicted % MGS values showing only moderate correlation. There were

exceptions to this, namely the standard descriptors, SOAPs, and hydration indices models, which made predictions with good correlation with the experimental values for the *Amino* dataset (PCC $\geq$ 0.60). Overall, these results were expected given the relatively high degree of uncertainty associated with the "splat cooling" assay, which often produces significant variation between measurements, as shown by the error distribution in Figure 1a. This uncertainty limits the maximum expected correlation and minimum expected error of the model to approximately 0.90 and 200, respectively (see Supplementary Figure 2 and Supplementary Information). The impact of this uncertainty is reduced in classification models, as the % MGS values are converted into categorical values. Although, as previously discussed, this also results in a loss of information and is susceptible to bias.

Given the success of the classification ensembles, we again explored an ensemble approach by obtaining the mean prediction across different combinations of three or more individual regression models. In this case, ensemble models also provide a means to quantify the uncertainty associated with each prediction (e.g. as the standard deviation (SD) of all the predictions). We ranked these models based on their PCC values, with the top three models and their components reported in Table 3. Similar to the ensemble classifiers, this approach improved the predictive capacity with respect to compounds in each dataset. This was particularly significant in the case of the *Amino* dataset, with the ensemble model generating test predictions with a correlation of 0.72. Whereas the best performing classification ensembles took a consensus across five or six models, optimal results were achieved by regressor ensembles which took an average across three or four models. This reflects the fact that a numerical average is more susceptible to the effects of anomalous predictions compared to a majority-voting consensus. Similarly, taking the mean across many models can have the effect of "smoothing out" the predicted values, especially at the tails of the target distribution.

Figure 2 shows the predictions of our best ensemble models versus the experimental % MGS values for the *Amino*, *Glyco2* and *Combined* datasets. These data, along with results shown in Table 3, highlight that models trained and evaluated using the *Amino* dataset yielded predictions with significantly higher correlation and lower error with respect to measured values, compared to the *Glyco2* or *Combined* datasets. This was surprising given that the *Glyco2* dataset was significantly larger than *Amino*, therefore provided many more examples to learn from during training. For the *Combined* dataset, the poor performance is suggestive of structural disparity between the amino acid and carbohydrate molecules. Indeed, Tanimoto similarities computed for these two distinct sets of compounds revealed significant structural diversity (Supplementary Figure 3), while dimensionality reduction performed on the molecular descriptors used in this work showed that these classes of molecules span different regions in feature space (Supplementary Figure 4). However, it remains unclear why *Glyco2* models could not achieve scores on par with the *Amino*:

Table 3 Performance metrics for test predictions obtained from individual and ensemble regression models. For the ensemble models, tick marks indicate the descriptor models included in the ensemble.

	Model	Std. descriptors	Molecular cliques	H-wACSFs	SOAPs	Hyd. histograms	Hyd. indices	MSE	PCC
Glyco2 dataset	Standard descriptors	✓						631	0.385
	Molecular cliques		✓					699	0.387
	H-wACSFs			✓				579	0.441
	SOAPs				✓			719	0.247
	Hydration histograms					✓		682	0.326
	Hydration indices[b]						✓	666	0.235
	Ensemble 1	✓	✓	✓				576	0.461
	Ensemble 2	✓	✓	✓	✓			575	0.457
	Ensemble 3		✓	✓	✓			576	0.456
Amino dataset	Standard descriptors	✓						673	0.590
	Molecular cliques		✓					966	0.451
	H-wACSFs			✓				738	0.535
	SOAPs				✓			624	0.639
	Hydration histograms					✓		960	0.366
	Hydration indices[b]						✓	667	0.625
	Ensemble 1	✓			✓		✓	503	0.716
	Ensemble 2			✓	✓		✓	509	0.713
	Ensemble 3		✓	✓	✓		✓	521	0.707
Combined dataset	Standard descriptors	✓						642	0.427
	Molecular cliques		✓					691	0.447
	H-wACSFs			✓				644	0.450
	SOAPs				✓			687	0.415
	Hydration histograms					✓		760	0.366
	Hydration indices[b]						✓	775	0.218
	Ensemble 1	✓	✓	✓				561	0.528
	Ensemble 2	✓	✓	✓		✓		570	0.520
	Ensemble 3		✓	✓	✓			581	0.520

[a]Multiple hydration indices with different hydration numbers used.

possible explanations include experimental inconsistencies within this literature dataset, or that the descriptors fail to capture important structural features for this class of molecule. Overall, each descriptor performed similarly across the different datasets, with a few exceptions such as the symmetry functions and hydration histograms, which gave significantly better predictions when used in conjunction with the *Glyco2* dataset (Table 3). Interestingly, a retrospective analysis also reveals that an ensemble approach does not offer the same improvements in performance relative to the individual models for the

Glyco2 dataset compared to the *Amino* dataset, perhaps for the same reasons noted previously.

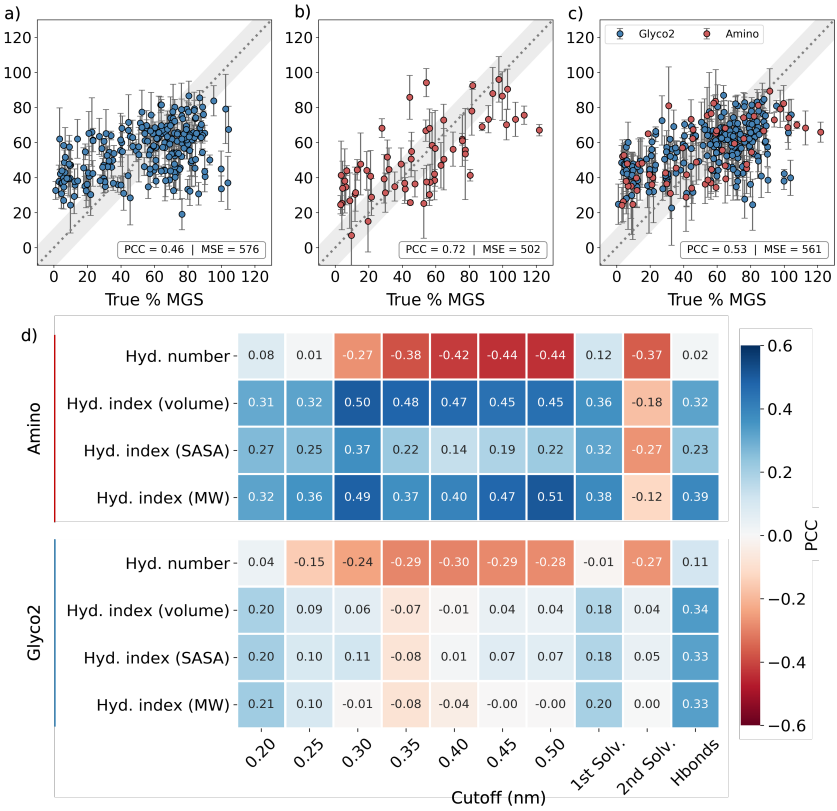

Fig. 2 Scatter plots showing the % MGS predictions from the best ensemble models trained using: a) the *Glyco* dataset; b) the *Amino* dataset; and c) the *Combined* dataset. Descriptors comprising the models are shown in Table 2. Error bars represent the ± one standard deviation for the individual predictions comprising the ensemble. Grey shaded region represents the typical error associated with experimental % MGS measurements, taken as the average standard deviation calculated across three measurements for all compounds in the *Amino* dataset. d) Heatmap showing the correlation (PCC) between % MGS values and hydration numbers or hydration indices calculated using different cutoffs and size metrics for the *Amino* and *Glyco2* datasets. Hydration indices were obtained by dividing the hydration numbers by either the molecular volume, SASA or molecular weight (MW), as indicated.

2.3 Hydration descriptors

The most striking results for any individual descriptor or model were the hydration indices, which achieved good performance for the *Amino* dataset (PCC = 0.63), despite performing poorly when used in conjunction with the *Glyco* or *Combined* dataset. This descriptor consists of ten hydration indices, where

each index represents a compound's hydration number normalised against its molecular volume. Hydration indices were first introduced by Tam and coworkers to explain the hydration-dependent IRI activity observed in carbohydrates, whereby more hydrated molecules showed greater levels of inhibition [21]. In the original formulation, hydration numbers were derived from molar compressibility coefficients according to the Passynsky equation [22], while the molar volume was calculated from density measurements [23]. In this work, hydration numbers were determined by finding the average number of water molecules within a given cutoff distance from the inhibitor molecule through a MD simulation trajectory, as described in the Methods. This quantity provides an estimate of the number of waters associated with each inhibitor, with a cutoff distance that can be tuned to capture information about the system at different ranges. To account for the molecule's size, we then normalised this value against different quantities (e.g. volume or surface area) also computed from MD simulations (see Methods), instead of molar volume which was used in the original work [21]. Hydration numbers were calculated using a number of fixed cutoffs as well the distance corresponding to the minima of the first and second solvation shells. These distances are determined from the hydration histograms illustrated in Supplementary Figure 13. The latter cutoffs are tailored to each molecule individually, giving hydration numbers which approximate the number of water molecules in the first and second solvation shells surrounding the solute, respectively. We also included the number of hydrogen-bonded waters as an alternative for the hydration number.

As expected, using a greater cutoff distance produced a greater hydration number and index for a given compound, as well as a broader distribution of these values for each dataset collectively (Supplementary Figure 5). We also found that the different hydration numbers and indices are correlated to varying degrees with IRI activity (% MGS), and that these trends were not shared between the *Amino* and *Glyco2* datasets. This data is summarised in Figure 2d. We observed negative correlations between the hydration numbers and % MGS values for both datasets when a distance cutoff greater than 0.3 nm was used, which includes the distance corresponding to the second solvation shell (Figure 2d). Yet, when these numbers were normalised against different size metrics to yield hydration indices, we found positive correlations with the *Amino* dataset, and few correlations for *Glyco2* (Figure 2d). For the *Amino* dataset, a positive correlation was observed for nearly all the different hydration indices, producing the strongest correlation when normalised by the molecular volume or weight. Meanwhile, only the indices calculated using hydrogen bonding data showed any correlation for the *Glyco2* dataset, following the same positive trend. The correlations between IRI activity and size metrics individually are provided in the Supplementary Information. We therefore found the opposite relationship to Tam and colleagues, who showed that the hydration indices for a set of nine mono- and disaccharides gave a strong negative correlation with MGS measurements [21]. This finding gave support

to a hypothesised IRI mechanism, wherein compounds with greater hydration (indices) cause more disruption to the ordering of surrounding QLL/bulk water, increasing the energy associated with the transfer of bulk water to a growing ice crystal via the QLL and consequently slowing ice growth [21]. Our computational hydration indices, both for this same set of nine sugars, as well as the entire *Glyco2* and *Amino* datasets, did not reproduce this correlation, irrespective of the cutoff distance used; instead we observed a moderate positive correlation with % MGS values (PCC ≈ 0.5) for the *Amino* dataset, and limited correlation for the *Glyco* dataset (PCC ≈ 0.3). Note that a positive correlation indicates that more "hydrated" molecules display weaker IRI.

Although our results are not in agreement with previous findings, these correlations do explain why the hydration indices were an effective ML descriptor for the amino acids, but not the carbohydrates. Moreover, is important to emphasise that whilst our computational hydration index is inspired by the previous work of Tam and colleagues [21], the values obtained using our method are not equivalent to those calculated using experimental data. Indeed, when the computational and experimental indices were compared for the set of mono- and disaccharides examined by Tam et al. [21], they showed only weak correlation (Supporting Figure 6). However, in light of limited experimental molar volume and compressibility coefficient data, a computational approach allows a hydration index to be calculated for virtually any chemical structure. Our analysis therefore included over 280 data points, encompassing highly active to inactive materials, compared to nine relatively inactive carbohydrates investigated in the aforementioned work. The size of these samples makes it challenging to draw comparative conclusions, as it is possible the previously reported trend may not hold for a larger set of small molecules, especially more active inhibitors. Overall, our computational hydration numbers and indices, derived in a different fashion but using similar intuition to the experimental properties, suggest a different interpretation. While it is not possible draw a causal relationship between hydration parameters and the mechanism of IRI, these results challenge the notion that hydration is a robust correlate of IRI for small molecules, and call for a review of this property and the hydration hypothesis.

2.4 Discovering novel amino acid IRIs

Returning to the results of our regression models, we found that the best predictions were obtained using an ensemble representing the mean % MGS prediction from three models (standard descriptors, SOAPs and hydration indices) trained and evaluated using the *Amino* dataset (Table 2). The % MGS predictions obtained using this model are shown in Figure 2b, achieving a MSE of 502 and PCC of 0.72. Encouraged by these results, we sought to use our model to predict the IRI activity for amino acids which had not yet been tested and thus represent novel small molecule inhibitors. This first required a dataset of new compounds for which % MGS predictions could be obtained. To achieve this, we assembled a prediction library consisting of around 500

amino acids, or amino acid-like compounds, as outlined in the Methods. These structures were all commercially available and therefore no chemical synthesis was required, which is a significant and deliberate advantage of using this amino acid platform to discover new IRI-active materials.

To obtain an % MGS prediction for each compound in this set, we again used an ensemble approach, taking the mean prediction across multiple models each trained on the entire *Amino* dataset. Given that the hydration indices are calculated from an MD trajectory, we opted to use an alternative ensemble model combining the standard descriptors, molecular cliques and SOAPs, which gave predictions with similar correlation (PCC = 0.67) and error (MSE = 570) for the test set, when evaluated using the same LOO CV approach. As these descriptors can be computed directly from the SMILES code for each molecule, this ensemble model can be easily deployed to screen the entire prediction library at minimal computational cost.

To build this ensemble model, % MGS predictions were obtained for each molecule in the prediction library using each of these descriptors, as outlined in the Methods. The SD of the predictions from individual descriptor models was also calculated to estimate the uncertainty associated with the ensemble predictions. Compounds with a prediction SD greater than 20 were subsequently removed from the library. The distributions of predicted % MGS values and associated error are shown in Supplementary Figure 7. These predictions were then ranked based on their % MGS values, and the ten most and ten least active molecules were designated as the prediction set, taking into account other practical considerations such as their cost and availability. Compounds with both low and high % MGS predictions were included in the prediction set to assess the model's accuracy in predicting IRI across a range of activities, despite the primary objective being to identify highly active inhibitors.

To verify these predictions experimentally, the compounds were then dissolved in 10 mM NaCl and tested at an initial concentration of 20 mM using the "splat cooling" assay. Of the 20 amino acids originally selected for the prediction set, the IRI activity was determined for 17 compounds, including three (compounds **7, 8** and **9**) which were tested at 10 mM as they were insoluble at the initial concentration. The predicted and experimental % MGS values for the prediction set are shown in Figure 3, alongside the chemical structures for a selection of compounds. Overall, the model achieved excellent performance, with 13 out of 17 predictions being in close agreement with the experimental % MGS, taking into account the uncertainty associated with the ensemble predictions and the measurement error. Altogether, the predictions achieved a PCC of 0.61 and MSE of 483 when compared against the measured values, which is a similar level of performance to that which was observed for the test set during cross-validation (Table 3).

For the molecules that were predicted to be highly active (% MGS < 20), three (**1, 3** and **5**) out of nine had only moderate IRI activity (% MGS ≈ 50) when tested experimentally at 20 mM (Figure 3a). These compounds all featured an aminopyridine scaffold with an O-methylated carboxylic acid (methyl

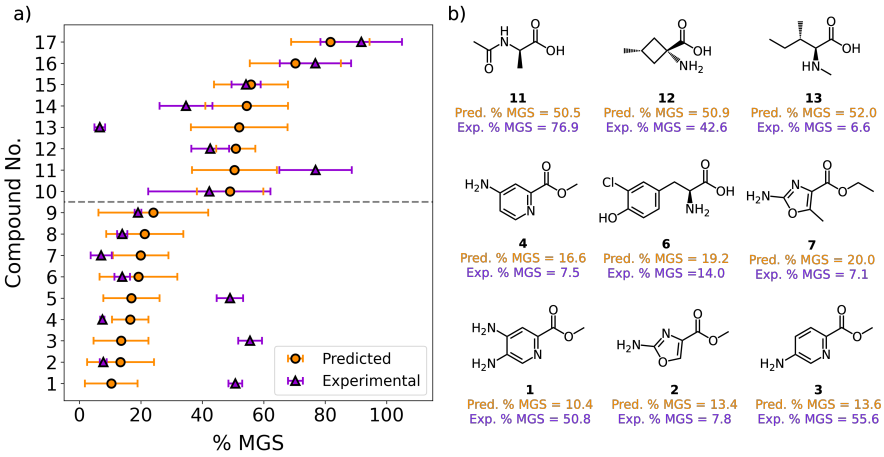

Fig. 3 a) Comparison of the predicted and experimental % MGS values for the 17 compounds in the prediction set. Compounds were tested at 20 mM in 10 mM NaCl, with the exception of compounds **7- 9**, which were tested at 10 mM. Error bars represent ± one standard deviation across three repeats. b) Chemical structures for a selection of compounds in the prediction set.

ester) group adjacent to the pyridine nitrogen (Figure 3a and Supplementary Figure 10). Interestingly, these same moieties are present in compound **4**, which was significantly more active than **1, 3** or **5**. The position of the amino group on the pyridine ring therefore appears crucial for IRI, with an amino group in the *para* position relative to the carboxylic acid/ester (e.g. in **1** and **3**) conferring only moderate activity. Other compounds correctly predicted to be highly active were chlorinated L-tyrosine derivatives **6** and **8**, nitropyridine carboxylic acid **9**, and aminooxazole esters **2** and **7** (Figure 3a and Supplementary Figure 10). In the case of the chlorinated tyrosines, it is interesting to note that IRI activity is retained with a hydroxyl group at the *para* position, unlike for hydrophilic *para*-amino and -cyano substitutions as discussed in our previous work [17]. It was suggested that the presence of these groups removes the hydrophobic face of the molecule – an essential motif for IRI – however in this case the chlorine atom(s) adjacent to the hydroxyl group in compounds **6** and **8** may have a mitigating effect. The most active hit among this set was the aminooxazole ester **7**, which was able to prevent growth almost entirely at 10 mM. Dose-dependency experiments revealed that this compound maintained IRI activity at 2.5 mM when tested in NaCl, which corresponds to less than 1 mg/mL, and was also active at 10 mM in PBS buffer (Supplementary Figures 14 and 15). Ice shaping assays performed using a nanolitre osmometer (see Supplementary Information) also confirmed that this compound does not alter the growth habit of individual ice crystals, suggesting no direct binding to the ice crystal surface (Supplementary Figure 16).

It is important to clarify that the IRI-active materials identified here do not represent entirely novel chemical scaffolds, bearing some resemblance to

molecules in the training set, which can be quantified by assessing their Tanimoto similarity (Supplementary Figure 9). We highlight that, rather than being an intrinsic limitation of the models, this is a product of the strategy used to compile the prediction set and library. First, the prediction library was generated via similarity search based on the 15 most active and 5 least active molecules among the training data. The compounds in the prediction library were then also filtered based on the uncertainty associated with the predictions (see Methods), which meant the prediction set was more likely to contain compounds with a high degree of structural similarity to molecules that the model has already "seen". Applying this model to a structurally distinct set of amino acids would make interesting work in the future. It is also worth acknowledging that the % MGS predictions for the active molecules were consistently higher than the experimental measurements, albeit still accurate within the associated prediction and measurement error. This is a consequence of averaging across many individual models, including those trained on different descriptors, to obtain an ensemble prediction. While individual descriptor models are capable of making predictions at the extremes of the % MGS distribution (i.e. close to 0 and 100 % MGS), the process of averaging means these values are less likely to be observed for ensemble predictions. Nonetheless, the ensemble approach increases the overall accuracy of the model, and the results presented here for the prediction set clearly justify this approach in retrospect.

Turning to the inactive predictions, the model also performed very well, with all predictions except two (compounds **11** and **13**) being accurate within the predicted margin of error. The chemical structures for these compounds are shown in Supplementary Figure 11. Compound **13** was N-methylated isoleucine (Figure 3b), which was surprisingly active given that N-modifications (methylation and/or acetylation) resulted in a loss of IRI activity for α-alanine and phenylalanine (Supplementary Figure 8), yet for isoleucine lowered the % MGS from 18.8 to 6.6. Similarly, while some of the compounds predicted to be inactive are structurally similar to molecules in the training set, we note that the cyclobutyl moiety found in compounds **10** and **12** represents a novel structure, highlighting the ability of the model to learn from training data and made extrapolations with good accuracy.

3 Conclusions

The discovery of materials that can inhibit ice recrystallisation is of fundamental importance for the development of cryoprotectants that can prevent freezing-induced damage to biological systems. Small molecules offer several advantages, such as biocompatibility, enhanced aqueous solubility and membrane permeability, which render them suitable for off-the-shelf applications in cryopreservation. However, the design of suitable IRI-active molecules has been a daunting task due to a limited knowledge of the molecular mechanisms underlying their activity, and the optimal structural features required for high inhibition. In this work, we present the first successful application of machine

learning to accurately predict the IRI activity of small molecules. Our approach leverages different molecular representations that incorporate both fine-grained 3D structural information and solvent interactions derived from MD simulations. These were applied to datasets compiled from literature, and a dataset obtained via our own experimental measurements; both of which have been made available to facilitate further research in this field. We employed this framework to perform a virtual screening of a commercial compound library and select novel IRI-active molecules with varying degrees of activity. Having verified these predictions experimentally, we reported the identification of a highly active aminooxazole ester **7** that can inhibit ice recrystallisation below 0.5 mg/mL, representing the most IRI-active amino acid discovered to date. We also investigated the role that molecular hydration plays in ice growth inhibition, revealing limited correlation between hydration numbers/indices and inhibition activity, which was previously offered as an mechanistic explanation for small molecule IRIs. Overall, our work highlights the power of data-driven approaches to facilitate the discovery of new cryoprotectants that can address the major challenge of freezing injury during cryopreservation. These molecules are crucial for unlocking the clinical potential of gene- and cell-based therapies, biomanufacturing, and cell and tissue banking.

4 Methods

4.1 Experimental measurements

Ice recrystallisation (inhibition) was measured using the "splat cooling" assay, as previously described by Knight et al. [24]. A 10 µL drop of each solution was dropped from a height of 1.4 m onto a glass coverslip placed on a thin aluminium plate cooled to −78 °C on dry ice. Upon impact with the coverslip, a polycrystalline ice wafer with an approximate diameter of 10 mm and thickness of 10 µm is formed instantly. The coverslip was then transferred to a Linkam Cryostage BCS196 pre-cooled to −8 °C and left for 30 minutes at −8 °C to anneal. Photographs were taken after 30 minutes via a Canon DSLR 500D digital camera using an Olympus CX 41 microscope equipped with a UIS-2 20x/0.45/∞/0-2/FN22 lens and crossed polarisers. The number of crystals in the field of view (FOV) were then counted using ImageJ [25], and this number was divided by the FOV area to give the mean grain size (MGS). The MGS for each sample was then compared to a positive control for ice growth, to obtain a percentage MGS (% MGS) relative to the control. Each experiment was performed in triplicate, and % MGS values were reported as the mean across the three repeats.

When performing this assay, it is essential to include saline (or other additives) in the solution to ensure liquid channels form between ice crystals. This allows for ice recrystallisation to occur and prevents false positive results [24]. Typically, using pure water and low concentrations of solutes, no ice growth occurs, generating a false positive. Considering this, a buffer such as phosphate-buffered saline (PBS) is typically employed (with ≈140 mM NaCl).

However, this concentration of saline is not essential and relevant IRI data has been reported in lower saline concentrations for other materials and small molecules [26, 27]. In this work, a 10 mM NaCl solution is used for all IRI activity measurements (i.e. all compounds in the *Amino* dataset). Compounds were assessed at 20 mM, unless otherwise stated.

4.2 Datasets

In this work, we have used three datasets broadly encompassing two classes of small molecule: amino acids and carbohydrates. The size and distribution of these datasets are shown in Table 1 and Figure 1. The *Glyco* and *Glyco2* datasets have been compiled using experimental data reported in peer-reviewed literature for a range of small molecule carbohydrates. The *Glyco* dataset is a subset of *Glyco2*, and was used in previous work to classify IRI-active small molecules [15]. In addition to the compounds in *Glyco*, *Glyco2* includes an additional 99 structures representing greater chemical diversity. This data was obtained from Refs [15, 21, 27–33]. The *Amino* dataset has been assembled via experimental measurements, as previously described in the Methods. Although the same experimental assay was used to obtain the data in *Glyco(2)* and *Amino* datasets, the different solution conditions used mean that values obtained are not always directly comparable. However, we emphasise that the aim of this work to predict IRI activity of molecules in saline, and a cross-comparison of % MGS values obtained under these different conditions (Supplementary Figure 12) suggests this approach to be valid.

To apply these models to identify novel IRI-active compounds, a set of unseen structures was also required. To obtain this set, we first took the 15 most active and 5 least active compounds in the *Amino* dataset and searched a catalogue of commercially available compounds (MolPort) for structures bearing similarity to the 20 selections. Similarity was determined by a Tanimoto coefficient [34] > 0.7, computed on based on an undefined descriptor via MolPort's web search tool. We chose to include both active and inactive structures in our similarity search to ensure greater coverage of chemical space, whilst providing the opportunity to validate the model's ability to accurately predict both high and low % MGS values. We note that while including only compounds with structural similarity to those in the training set does limit the capacity of the model to identify truly novel IRI-active scaffolds, given the limited number of training examples this approach ensures the prediction set is within the model's domain of applicability. The prediction set was then filtered to remove any compounds appearing in the training set (i.e. hard overlap), as well as compounds containing co-salts or co-additives, or with a predicted logP > 1.8 (i.e. likely insoluble), computed via RDKit [35]. The final prediction set comprised 497 unique structures. Consensus % MGS predictions for each compounds were then obtained as outlined below. Predictions with a high degree of uncertainty (SD > 15 % MGS) were removed, and the 10 most active and 10 least active compounds remaining were then purchased and tested experimentally, as described below.

4.3 Molecular descriptors

We have used six different descriptors in this work, encompassing low (0D) and high dimensional (4D) molecular representations. Four of these descriptors have been repurposed following success in a range of tasks such as chemical property prediction (e.g. enthalpies [36], potential energies [37], lipophilicity [38]) and molecular generation [39]:

- "Standard" descriptors – A collection of ~ 45 molecular properties that are accessible via RDKit [35], e.g. molecular weight, atom/bond counts, cLogP, topological polar surface area. A complete list of properties can be found in the GitHub repository [18].
- Molecular cliques – Considering a molecule as a collection of nodes (atoms) and edges (bonds), i.e. a 2D graph, a clique represents a subgraph of a molecule. A vocabulary of cliques is constructed for a given dataset, and the cliques for each molecule are encoded as a fingerprint. For more information, see Ref. [38].
- Histograms of weighted atom-centred symmetry functions (H-wACSF) – Symmetry functions describe the local (3D) chemical environment of an atom in a molecule using radial and angular symmetry functions based on the distance and angles between pairs and triplets of atoms, respectively. In this formulation, element-dependent weighted symmetry functions are computed and the values then binned to obtain a histogram-like descriptor with the same (reduced) dimensionality for all molecules, independent of their size and atomic composition. H-wACSF parameters used here are listed in Supplementary Information; for more detail, see Ref. [38].
- Smooth overlap of atomic positions (SOAP) descriptor – 3D atomic environments encoded using atomic density fields composed of Gaussian functions centred on each atom. This formalism is extended to describe molecules by averaging the density field across the constituent atoms. SOAP parameters were optimised using a genetic algorithm, as described in Ref. [40]. The optimal parameter set is provided in Supplementary Information, and readers are directed to Refs [37, 40] for more information.

We have also engineered two new and relatively simple descriptors bespoke for use case via MD simulations:

- Hydration histograms – From short MD simulations, we compute a probability density histogram by calculating the pairwise distances between each water molecule and the solute from configurations sampled from the trajectory. These distances d are then binned and normalised based on the total number of distances considered, and the interval width Δd, resulting in probability densities $P(d)$ for each bin, given by $P(d) = \frac{n_d}{\sum_i^{D_{\text{cut}}} n_i \Delta d}$. Histograms were computed here using 100 bins, up to $d_{\text{cut}} = 0.5$ nm, hence $\Delta d = 0.005$ nm.
- Hydration indices – First defined by Tam and colleagues [21], a hydration index represents the number of nonexchangeble water molecules associated

with a solute's "hydration layer" (i.e. hydration number), divided by its partial molar volume. Here, we compute the hydration index computationally by means of MD simulations. Our hydration numbers represent the numbers of waters hydrogen-bonded to the solute, determined via geometric criteria, as well as the numbers of water molecules within a given cutoff distance of the molecule. These values are normalised against the compound's molecular volume, computed via RDKit [35], yielding hydration indices. The hydration indices descriptor includes ten indices, where the hydration numbers were computed using seven fixed cutoff distances in the range 0.20 - 0.50 nm, the distances corresponding to the first and second solvation shells, and hydrogen bond numbers.

Excluding the standard descriptors and cliques, these representations were constructed from 3D atomic coordinates. To generate corresponding 3D conformations, short MD simulations of each system were performed as follows. Each compound was first solvated in water (TIP4P/Ice [41]) in a 4 nm cubic cell and simulated for 20 ns at 273 K via GROMACS 5.1.3 [42], using the CHARMM36 forcefield [43]. The final conformation was used to construct H-wACSFs and SOAPs, whereas the hydration descriptors were averaged over 100 different conformations sampled from the trajectories.

Given that many amino acids can exist in multiple ionisation states, it was important to model the correct form. To determine the predominant state of amino acids under neutral solution conditions, putative ionisation states between the range of pH 6.5 and 7.5 were predicted using the Diamorphite-DL package [44] with a precision factor of 1.0. For structures with multiple predicted states, the final state was selected manually based on pK_a values for the compound found in literature or estimated via the MolGpKa tool [45]. The ionisation states of a random sample were then checked independently to verify the results of this procedure.

4.4 Classification models

All models were trained and evaluated using Keras (Tensorflow) [46], alongside scikit-learn [47]. To perform classification, numerical % MGS values were first converted via one-hot encoding using a defined threshold for activity. The descriptor features were also scaled between 0 and 1, using Min-Max scaling. Independent models were then trained using each of the six descriptors. A randomised grid search was performed in combination with manual tuning to identify the optimal hyperparameters for each model; these hyperparameters are reported in the Supplementary Information. A leave-one-out (LOO) cross validation (CV) procedure was to train and evaluate each model. 10 % of the training data was randomly selected and used as a validation set in conjunction with an early stopping criterion to prevent overfitting. Models were trained over a maximum of 300 epochs, using the binary cross entropy loss function. A classification threshold (probability) of 0.5 was used throughout. In cases where classes were imbalanced, the Synthetic Minority Over-sampling TEchnique

(SMOTE) was used to bootstrap the training set to achieve equal number of active/inactive observations. For ensemble classification models, predictions were combined and averaged via a majority voting scheme. In this scheme, the class with the highest number of votes is used. In cases where each class had an equal number of votes, predictions labels defaulted to inactive

4.5 Regression models

% MGS values and descriptor features were first scaled between 0 and 1 using Min-Max scaling. The hyperparameters for each model were identified using a randomised search grid followed by manual tuning, are are given in the Supplementary Information. The same LOO CV procedure described above for classification was also used to train and evaluate the regression models, unless otherwise stated. Models were trained over a maximum of 300 epochs, employing the mean squared error (MSE) as the loss function. L2 (ridge) regularisation with $\sigma = 0.005$ was also used to prevent overfitting when using certain descriptors (e.g. SOAPs). Ensemble or consensus predictions were calculated as the mean predicted value across a given set of individual descriptors, while the standard deviation was used to estimate the associated uncertainty.

Supplementary information. Additional experimental and computational details and methods including: nanoliter osmometry, classification metrics, molecular descriptor parameters, neural network hyperparameters, repeated LOO CV results, maximum expected performance estimation, Tanimoto similarities and dimensionality reduction. Supplementary figures including: Tanimoto coefficient distributions, dimensionality reduction biplots, hydration number and index distributions, % MGS prediction distribution, chemical structures for the prediction set, additional IRI measurements in NaCl and PBS, hydration descriptors schematic, and cryomicrographs from nanolitre osmometry.

Acknowledgements. M.T.W. thanks the MRC for a studentship through the MRC Doctoral Training Partnership in Interdisciplinary Biomedical Research (grant no. MR/S502534/1). M.I.G. thanks the ERC for a Consolidator Grant (866056) and the Royal Society for an Industry Fellowship (191037) joint with Cytivia. G.C.S. thanks the BBSRC for a Research Grant (grant no. BB/V015559/1). We gratefully acknowledge the use of the ARCHER2 UK National Supercomputing Service (https://www.archer2.ac.uk), which we have accessed via the HecBioSim consortium, funded by the EPSRC (grant no. EP/R029407/1). We also gratefully acknowledge the use of Athena at HPC Midlands+, which was funded by the UK Engineering and Physical Sciences Research Council (grant no. EP/P020232/1), via the HPC Midlands+ consortium. We would also like to acknowledge the high-performance computing facilities provided by the Scientific Computing Research Technology Platform at the University of Warwick.

Author contributions. M.T.W performed the computational work and analysis. M.T.W, C.I.B and A.B performed the experimental work and analysis. All authors interpreted the results. G.C.S. and M.I.G. conceived the research. M.T.W., G.C.S., and M.I.G. wrote the manuscript.

References

[1] Giwa, S., Lewis, J.K., Alvarez, L., Langer, R., Roth, A.E., Church, G.M., Markmann, J.F., Sachs, D.H., Chandraker, A., Wertheim, J.A., Rothblatt, M., Boyden, E.S., Eidbo, E., Lee, W.P.A., Pomahac, B., Brandacher, G., Weinstock, D.M., Elliott, G., Nelson, D., Acker, J.P., Uygun, K., Schmalz, B., Weegman, B.P., Tocchio, A., Fahy, G.M., Storey, K.B., Rubinsky, B., Bischof, J., Elliott, J.A.W., Woodruff, T.K., Morris, G.J., Demirci, U., Brockbank, K.G.M., Woods, E.J., Ben, R.N., Baust, J.G., Gao, D., Fuller, B., Rabin, Y., Kravitz, D.C., Taylor, M.J., Toner, M.: The promise of organ and tissue preservation to transform medicine. Nature Biotechnology **35**(6), 530–542 (2017). https://doi.org/10.1038/nbt.3889

[2] Meneghel, J., Kilbride, P., Morris, G.J.: Cryopreservation as a Key Element in the Successful Delivery of Cell-Based Therapies—A Review. Frontiers in Medicine **7**, 592242 (2020). https://doi.org/10.3389/fmed.2020.592242

[3] Fowler, A., Toner, M.: Cryo-injury and biopreservation. Annals of the New York Academy of Sciences **1066**(1), 119–135 (2005). https://doi.org/10.1196/annals.1363.010

[4] Costanzo, J.P., Lee, R.E.: Avoidance and tolerance of freezing in ectothermic vertebrates. Journal of Experimental Biology **216**(11), 1961–1967 (2013). https://doi.org/10.1242/jeb.070268

[5] Raymond, J.A., DeVries, A.L.: Adsorption inhibition as a mechanism of freezing resistance in polar fishes. Proceedings of the National Academy of Sciences of the United States of America **74**(6), 2589–2593 (1977). https://doi.org/10.1073/pnas.74.6.2589

[6] Liu, S., Ben, R.N.: C-linked galactosyl serine AFGP analogues as potent recrystallization inhibitors. Organic Letters **7**(12), 2385–2388 (2005). https://doi.org/10.1021/ol050677x

[7] Biggs, C.I., Bailey, T.L., Ben Graham, Stubbs, C., Fayter, A., Gibson, M.I.: Polymer mimics of biomacromolecular antifreezes. Nature Communications **8** (2017). https://doi.org/10.1038/s41467-017-01421-7

[8] Drori, R., Li, C., Hu, C., Raiteri, P., Rohl, A.L., Ward, M.D., Kahr, B.: A Supramolecular Ice Growth Inhibitor. Journal of the American Chemical Society **138**(40), 13396–13401 (2016). https://doi.org/10.1021/jacs.

6b08267

[9] Capicciotti, C.J., Kurach, J.D.R., Turner, T.R., Mancini, R.S., Acker, J.P., Ben, R.N.: Small molecule ice recrystallization inhibitors enable freezing of human red blood cells with reduced glycerol concentrations. Scientific Reports **5** (2015). https://doi.org/10.1038/srep09692

[10] Briard, J.G., Jahan, S., Chandran, P., Allan, D., Pineault, N., Ben, R.N.: Small-Molecule Ice Recrystallization Inhibitors Improve the Post-Thaw Function of Hematopoietic Stem and Progenitor Cells. ACS Omega **1**(5), 1010–1018 (2016). https://doi.org/10.1021/acsomega.6b00178

[11] Lautner, L., Himmat, S., Acker, J.P., Nagendran, J.: The efficacy of ice recrystallization inhibitors in rat lung cryopreservation using a low cost technique for ex vivo subnormothermic lung perfusion. Cryobiology **97**, 93–100 (2020). https://doi.org/10.1016/j.cryobiol.2020.10.001

[12] Galvao, J., Davis, B., Tilley, M., Normando, E., Duchen, M.R., Cordeiro, M.F.: Unexpected low-dose toxicity of the universal solvent dmso. The FASEB Journal **28**(3), 1317–1330 (2014)

[13] Verheijen, M., Lienhard, M., Schrooders, Y., Clayton, O., Nudischer, R., Boerno, S., Timmermann, B., Selevsek, N., Schlapbach, R., Gmuender, H., Gotta, S., Geraedts, J., Herwig, R., Kleinjans, J., Caiment, F.: DMSO induces drastic changes in human cellular processes and epigenetic landscape in vitro. Scientific Reports **9**(1), 1–12 (2019). https://doi.org/10.1038/s41598-019-40660-0

[14] Elliott, G.D., Wang, S., Fuller, B.J.: Cryoprotectants: A review of the actions and applications of cryoprotective solutes that modulate cell recovery from ultra-low temperatures. Cryobiology **76**, 74–91 (2017). https://doi.org/10.1016/j.cryobiol.2017.04.004

[15] Briard, J.G., Fernandez, M., De Luna, P., Woo, T.K., Ben, R.N.: QSAR Accelerated Discovery of Potent Ice Recrystallization Inhibitors. Scientific Reports **6**(1), 26403 (2016)

[16] Warren, M.T., Galpin, I., Bachtiger, F., Gibson, M.I., Sosso, G.C.: Ice Recrystallization Inhibition by Amino Acids: The Curious Case of Alpha- and Beta-Alanine. The Journal of Physical Chemistry Letters **13**, 2237–2244 (2022). https://doi.org/10.1021/acs.jpclett.1c04080

[17] Warren, M.T., Galpin, I., Hasan, M., Hindmarsh, S.A., Padrnos, J.D., Edwards-Gayle, C., Mathers, R.T., Adams, D.J., Sosso, G.C., Gibson, M.I.: Minimalistic ice recrystallisation inhibitors based on phenylalanine. Chemical Communications **58**(55), 7658–7661 (2022). https://doi.org/10.1039/d2cc02531k

[18] Sosso, G.C.: Data-driven discovery of potent ice recrystallisation inhibitors. https://github.com/gcsosso/DOLMEN. GitHub (2023)

[19] Ho, T.K.: Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282 (1995). IEEE

[20] Ericksen, S.S., Wu, H., Zhang, H., Michael, L.A., Newton, M.A., Hoffmann, F.M., Wildman, S.A.: Machine learning consensus scoring improves performance across targets in structure-based virtual screening. Journal of chemical information and modeling **57**(7), 1579–1590 (2017)

[21] Tam, R.Y., Ferreira, S.S., Czechura, P., Ben, R.N., Chaytor, J.L.: Hydration index-a better parameter for explaining small molecule hydration in inhibition of ice recrystallization. Journal of the American Chemical Society **130**(51), 17494–17501 (2008). https://doi.org/10.1021/ja806284x

[22] Galema, S.A., Høiland, H.: Stereochemical aspects of hydration of carbohydrates in aqueous solutions. 3. Density and ultrasound measurements. Journal of Physical Chemistry **95**(13), 5321–5326 (1991). https://doi.org/10.1021/j100166a073

[23] Høiland, H., Holvik, H.: Partial molal volumes and compressibilities of carbohydrates in water. Journal of Solution Chemistry **7**(8), 587–596 (1978). https://doi.org/10.1007/BF00646036

[24] Knight, C.A., Hallett, J., DeVries, A.L.: Solute effects on ice recrystallization: An assessment technique. Cryobiology **25**(1), 55–60 (1988). https://doi.org/10.1016/0011-2240(88)90020-X

[25] Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.Y., White, D.J., Hartenstein, V., Eliceiri, K., Tomancak, P., Cardona, A.: Fiji: An open-source platform for biological-image analysis. Nature Methods **9**(7), 676–682 (2012). https://doi.org/10.1038/nmeth.2019

[26] Georgiou, P.G., Marton, H.L., Baker, A.N., Congdon, T.R., Whale, T.F., Gibson, M.I.: Polymer Self-Assembly Induced Enhancement of Ice Recrystallization Inhibition. Journal of the American Chemical Society **143**(19), 7449–7461 (2021). https://doi.org/10.1021/jacs.1c01963

[27] Balcerzak, A.K., Febbraro, M., Ben, R.N.: The importance of hydrophobic moieties in ice recrystallization inhibitors. RSC Advances **3**(10), 3232–3236 (2013). https://doi.org/10.1039/c3ra23220d

[28] Jackman, J., Noestheden, M., Moffat, D., Pezacki, J.P., Findlay, S., Ben, R.N.: Assessing antifreeze activity of AFGP 8 using domain recognition

software. Biochemical and Biophysical Research Communications **354**(2), 340–344 (2007). https://doi.org/10.1016/j.bbrc.2006.12.225

[29] Balcerzak, A.K., Ferreira, S.S., Trant, J.F., Ben, R.N.: Structurally diverse disaccharide analogs of antifreeze glycoproteins and their ability to inhibit ice recrystallization. Bioorganic and Medicinal Chemistry Letters **22**(4), 1719–1721 (2012). https://doi.org/10.1016/j.bmcl.2011.12.097

[30] Capicciotti, C.J., Leclère, M., Perras, F.A., Bryce, D.L., Paulin, H., Harden, J., Liu, Y., Ben, R.N.: Potent inhibition of ice recrystallization by low molecular weight carbohydrate-based surfactants and hydrogelators. Chemical Science **3**(5), 1408–1416 (2012). https://doi.org/10.1039/c2sc00885h

[31] Capicciotti, C.J., Mancini, R.S., Turner, T.R., Koyama, T., Alteen, M.G., Doshi, M., Inada, T., Acker, J.P., Ben, R.N.: O-Aryl-Glycoside Ice Recrystallization Inhibitors as Novel Cryoprotectants: A Structure-Function Study. ACS Omega **1**(4), 656–662 (2016). https://doi.org/10.1021/acsomega.6b00163

[32] Trant, J.F., Biggs, R.A., Capicciotti, C.J., Ben, R.N.: Developing highly active small molecule ice recrystallization inhibitors based upon C-linked antifreeze glycoprotein analogues. RSC Advances **3**(48), 26005–26009 (2013). https://doi.org/10.1039/c3ra43835j

[33] Briard, J.G., Jahan, S., Chandran, P., Allan, D., Pineault, N., Ben, R.N.: Small-Molecule Ice Recrystallization Inhibitors Improve the Post-Thaw Function of Hematopoietic Stem and Progenitor Cells. ACS Omega **1**(5), 1010–1018 (2016). https://doi.org/10.1021/acsomega.6b00178

[34] Bajusz, D., Rácz, A., Héberger, K.: Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of cheminformatics **7**(1), 1–13 (2015)

[35] Landrum, G.: Rdkit: Open-source cheminformatics software (2016)

[36] Gastegger, M., Schwiedrzik, L., Bittermann, M., Berzsenyi, F., Marquetand, P.: WACSF - Weighted atom-centered symmetry functions as descriptors in machine learning potentials. Journal of Chemical Physics **148**(24) (2018) https://arxiv.org/abs/1712.05861v1. https://doi.org/10.1063/1.5019667

[37] Bartók, A.P., Kondor, R., Csányi, G.: On representing chemical environments. Physical Review B **87**(18) (2013). https://doi.org/10.1103/physrevb.87.184115

[38] Barnard, T., Hagan, H., Tseng, S., Sosso, G.C.: Less may be more: An

informed reflection on molecular descriptors for drug design and discovery. Molecular Systems Design and Engineering **5**(1), 317–329 (2020). https://doi.org/10.1039/c9me00109c

[39] Jin, W., Barzilay, R., Jaakkola, T.: Junction tree variational autoencoder for molecular graph generation. arXiv:1802.04364 **5**, 3632–3648 (2018)

[40] Barnard, T., Tseng, S., Darby, J.P., Bartók, A.P., Broo, A., Sosso, G.C.: Leveraging genetic algorithms to maximise the predictive capabilities of the SOAP descriptor. Molecular Systems Design & Engineering (2022). https://doi.org/10.1039/d2me00149g

[41] Abascal, J.L.F., Sanz, E., Fernández, R.G., Vega, C.: A potential model for the study of ices and amorphous water : TIP4P / Ice. The Journal of Chemical Physics **122** (2005)

[42] Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., Lindah, E.: Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX **1-2**, 19–25 (2015). https://doi.org/10.1016/j.softx.2015.06.001

[43] Guvench, O., Mallajosyula, S.S., Raman, E.P., Hatcher, E., Vanommeslaeghe, K., Foster, T.J., Jamison, F.W., Mackerell, A.D.: CHARMM additive all-atom force field for carbohydrate derivatives and its utility in polysaccharide and carbohydrate- protein modeling. Journal of Chemical Theory and Computation **7**, 3162–3180 (2011). https://doi.org/10.1021/ct200328

[44] Ropp, P.J., Kaminsky, J.C., Yablonski, S., Durrant, J.D.: Dimorphite-DL: An open-source program for enumerating the ionization states of drug-like small molecules. Journal of Cheminformatics **11**(1), 1–8 (2019). https://doi.org/10.1186/s13321-019-0336-9

[45] Pan, X., Wang, H., Li, C., Zhang, J.Z.H., Ji, C.: MolGpka: A Web Server for Small Molecule pKaPrediction Using a Graph-Convolutional Neural Network. Journal of Chemical Information and Modeling **61**(7), 3159–3165 (2021). https://doi.org/10.1021/acs.jcim.1c00075

[46] Chollet, F.: Keras. GitHub (2015)

[47] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Édouard Duchesnay: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)