Green Chemistry

PAPER

View Article Online
View Journal | View Issue

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Cite this: *Green Chem.*, 2016, **18**, 4461

Probabilistic diagram for designing chemicals with reduced potency to incur cytotoxicity†‡

Longzhu Q. Shen,[a] Richard S. Judson,[b] Fjodor Melnikov,[a] John Roethle,[c] Aditya Gudibanda,[d] Julie B. Zimmerman[a] and Paul T. Anastas*[a]

Received 14th April 2016,
Accepted 5th July 2016

DOI: 10.1039/c6gc01058j

www.rsc.org/greenchem

Toxicity is a concern with many chemicals currently in commerce, and with new chemicals that are introduced each year. The standard approach to testing chemicals is to run studies in laboratory animals (*e.g.* rats, mice, dogs), but because of the expense of these studies and concerns for animal welfare, few chemicals besides pharmaceuticals and pesticides are fully tested. Over the last decade there have been significant developments in the field of computational toxicology which combines *in vitro* tests and computational models. The ultimate goal of this field is to test all chemicals in a rapid, cost effective manner with minimal use of animals. One of the simplest measures of toxicity is provided by high-throughput *in vitro* cytotoxicity assays, which measure the concentration of a chemical that kills particular types of cells. Chemicals that are cytotoxic at low concentrations tend to be more toxic to animals than chemicals that are less cytotoxic. We employed molecular characteristics derived from density functional theory (DFT) and predicted values of log(octanol–water partition coefficient) (log *P*) to construct a design variable space, and built a predictive model for cytotoxicity based on U.S. EPA Toxicity ForeCaster (ToxCast) data tested up to 100 μM using a Näive Bayesian algorithm. External evaluation showed that the area under the curve (AUC) for the receiver operating characteristic (ROC) of the model to be 0.81. Using this model, we provide probabilistic design rules to help synthetic chemists minimize the chance that a newly synthesized chemical will be cytotoxic.

1. Introduction

Despite the tremendous benefits of modern, man-made chemicals and the products they go into, some of these chemicals possess unintended biological activities that pose a threat to public health and the environment. In order to reduce the chance of undesirable health effects induced by chemicals, one can carry out toxicology studies, traditionally using laboratory animals such as rats, mice or dogs. However, these studies are expensive (millions of dollars per chemical) and require the sacrifice of large numbers of animals. As a result, many chemicals are put on the market with little to no toxicity testing.[1] It is estimated that about 83% of chemicals in commerce lack safety data.[2] This has motivated the development of new *in vitro* and *in silico* methods to evaluate chemical toxicity and safety. Ideally, a new approach would allow all chemicals to be adequately tested, and do so at a reasonable cost and with minimal use of experimental animals.

In vitro high throughput screening (HTS) methods have emerged as an efficient technology to examine how chemicals disrupt biological pathways and lead to adverse health outcomes. A paradigm shift from *in vivo* to *in vitro* and *in silico* testing was described by the U.S. National Research Council (NRC) in their report on Toxicity Testing in the 21st Century.[3] In order to evaluate practical approaches to implement the "Tox21" vision, U.S. National Toxicological Program (NTP), the U.S. Environmental Protection Agency (EPA) and the NIH National Center for Advancing Translational Sciences (NCATS) collaboratively forged a research partnership. This Tox21 partnership is using HTS methods to test thousands of chemicals in a wide variety of cells, pathways and technologies, relevant to many aspects of chemical toxicity.[4–9]

Large data sets (thousands of chemicals, hundreds of measurements per chemical) are ideal for developing machine learning predictive models.[10,11] While the Tox21 collaborators are primarily focused on predicting toxicity of largely-untested existing chemicals,[12] these data sets can also be used in the area of green chemistry to help develop rules to apply in the

[a]*School of Forestry and Environmental Studies, New Haven, CT, 06511, USA. E-mail: paul.anastas@yale.edu; Fax: +1-203-436-8574; Tel: +1-203-432-5215*
[b]*U.S. EPA, National Center for Computational Toxicology, RTP NC 27711, USA*
[c]*Department of Chemistry, Yale University, New Haven, CT 06511, USA*
[d]*Department of Computer Science, Yale University, New Haven, CT 06511, USA*

† EPA disclaimer: the views expressed in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

‡ Electronic supplementary information (ESI) available. See DOI: 10.1039/c6gc01058j

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

design of newer, safer chemicals.[13] Owing to the difference in research goals, the two fields apply separate biases in how models are developed. For computational toxicology, high predictive power for toxicity is the paramount object. A model can be as complex as is needed to achieve high accuracy, as long as the training data is not overfitted. The physical meanings of the input variables or model descriptors are often of secondary concerns. In contrast, green design[13] pays heavy attention to the physical meanings of the computed molecular descriptors, their numerical accuracy, and the ability to map them to variables which synthetic chemists have control. The variables in the toxicity model are to be used as yardsticks for chemists in designing molecules. Physical observables and chemically intuitive variables are much preferable to more abstract global descriptions of chemical structure. A sufficiently accurate toxicity model that includes easy-to-interpret control variables will greatly aid in the green design process.

Current examples of green design research have modeled acute aquatic toxicity[14,15] and mutagenicity/carcinogenicity[16] endpoints. In this current report, we focus on *in vitro* cytotoxicity as a good testing ground for the development of green design rules associated with chemical toxicity. First, the data sets for cytotoxicity are larger, more uniform in protocol, and less noisy than any existing *in vivo* toxicity data sets. Second, cytotoxicity is on its own a valuable measure for safety assessment. It is normally performed in the early stage of drug development as a first filter.[17] It can also, to certain degree, serve as an indicator of the range of doses where one might see *in vivo* toxicity.[18,19] There exists high research interest in predictive modeling for the cytotoxicity endpoint.[20–24] Therefore, it would be useful to be able to design commercial chemicals with reduced cytotoxicity potency. (Note that all chemicals are cytotoxic, and what matters is the concentration or dose that is required to cause it.)

There are three main challenging tasks in green design in the context of toxicity. The first one is to select and generate physically meaningful design variables that have an effect on the underlying mechanisms responsible for the toxicity endpoint of interest. The second one is to construct a sufficiently predictive model of toxicity based on these design variables. The third one is to devise a method to present the design-variable/toxicity mapping in a way that provides an easy way to guide safer molecular design. This current project addresses all three of these issues in the context of a particular mode of toxicity, namely cytotoxicity. We produce a probabilistic design diagram as a methodological addition to existing molecular design tools. This diagram renders the convenience of searching for solutions in the physical property space for a customerized likelihood of not causing cytotoxicity in human cells (named as benign probability hereafter). This diagram, we believe, offers advantages in its use for guiding the design of safer chemicals.

2. Methods

2.1. Cytotoxicity data resource

U.S. EPA Toxicity ForeCaster (ToxCast) program[25] phase I and II chemical library contains a diverse collection of chemicals profiled across 821 *in vitro* endpoints. Among these assays, we chose 37 that are related to cytotoxicity in this study, listed in Table 1.

2.2. Chemical selection

Chemicals were initially selected from the ToxCast library if they had been tested in more than 90% of the assays in Table 1. (Note that these chemicals have also been tested in the majority of assays in the overall ToxCast program: see Fig. S1 in ESI‡). Within this group of chemicals, three criteria were set to select candidates for modeling cytotoxicity.

1. Single compound with a definite structure, excluding geometrical and optical isomers and mixtures;
2. Containing no metal elements;
3. Molecular weight < 1000.

A total of 1006 chemicals met these criteria. These chemicals were further broken down into two classes. A chemical tested to be positive in two or more cytotoxicity assays (Table 1) was labeled as "active".[26] Otherwise it was designated as "inactive". The chemical hit rates distribution is provided in Fig. S2 in ESI.‡ Note that in these assays, we typically tested up to 100 μM, so cytotoxicity at higher concentrations would not be seen. Therefore, we are strictly modeling "cytotoxicity below 100 μM". This classification strategy resulted in two balanced classes with a membership ratio of 0.96. All chemical data were evenly split into a training and testing set for the purpose of model cross-validation and external evaluation. Selected chemicals were first desalted using the open source chemistry toolbox OpenBabel.[27] Then 3D structures with the lowest energy conformer were generated using ChemAxon Marvin calculator plugins.[28]

2.3. Generation and selection of design variables

Six design variables were employed in this study: molecular softness (SOF), electrophilicity index (EPH), ionization potential (IP), electron affinity in the aqueous phase (EA.aq), polarizability (PLRZ) and $\log P$. These variables are physically meaningful and chemically intuitive. The mapping between these variables and the variables over which the chemists have direct control in synthesis is an important topic in green chemistry. $\log P$ was calculated using the ChemAxon Marvin calculator plugin.[28] The remaining variables were computed based on DFT implemented in Gaussian 09 rev.D.01.[29] Boese and Martin's τ-dependent hybrid functional[30] and basis set 6-31+G(d) were used for full geometry optimization. Vertical IP and electron affinity (EA) were calculated in vacuum. The species with an extra electron were then transferred into the implicit aqueous environment based on the universal solvation model[31] to obtain vertical EA.aq. SOF and EPH were derived according the following formulas.§[32]

§ Numerical values of design variables used in the study are available upon request.

 This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Table 1 Description for cytotoxicity-related assays

Assay name	Biological process	Organism	Tissue	Cell type
ACEA_T47D_80 h_negative	Cell proliferation	Human	Breast	T47D
APR_HepG2_CellLoss_24 h_dn	Cell death	Human	Liver	HepG2
APR_HepG2_CellLoss_72 h_dn	Cell death	Human	Liver	HepG2
BSK_3C_proliferation_down	Cell proliferation	Human	Vascular	Umbilical vein endothelium
BSK_3C_SRB_down	Cell death	Human	Vascular	Umbilical vein endothelium
BSK_3C_Vis_down	Cell morphology	Human	Vascular	Umbilical vein endothelium
BSK_4H_SRB_down	Cell death	Human	Vascular	Umbilical vein endothelium
BSK_BE3C_SRB_down	Cell death	Human	Lung	Bronchial epithelial cell
BSK_CASM3C_proliferation_down	Cell proliferation	Human	Vascular	Umbilical vein endothelium and coronary artery smooth muscle cells
BSK_CASM3C_SRB_down	Cell death	Human	Vascular	Umbilical vein endothelium and coronary artery smooth muscle cells
BSK_hDFCGF_proliferation_down	Cell proliferation	Human	Skin	Foreskin fibroblast
BSK_hDFCGF_SRB_down	Cell death	Human	Skin	Foreskin fibroblast
BSK_KF3CT_SRB_down	Cell death	Human	Skin	Keratinocytes and foreskin fibroblasts
BSK_LPS_SRB_down	Cell death	Human	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells
BSK_SAg_PBMCCytotoxicity_down	Cell death	Human	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells
BSK_SAg_proliferation_down	Cell proliferation	Human	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells
BSK_SAg_SRB_down	Cell death	Human	Vascular	Umbilical vein endothelium and peripheral blood mononuclear cells
Tox21_AR_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_ERa_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_GR_BLA_antagonist_viability	Cell proliferation	Human	Cervix	HeLa
Tox21_MitochondrialToxicity_viability	Cell proliferation	Human	Liver	HepG2
Tox21_FXR_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_PPARd_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_PPARg_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293
Tox21_VDR_BLA_antagonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_ARE_BLA_agonist_viability	Cell proliferation	Human	Liver	HepG2
Tox21_HSE_BLA_agonist_viability	Cell proliferation	Human	Cervix	HeLa
Tox21_p53_BLA_p1_viability	Cell proliferation	Human	Intestinal	HCT116
Tox21_FXR_BLA_agonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_PPARd_BLA_agonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_p53_BLA_p2_viability	Cell proliferation	Human	Intestinal	HCT116
Tox21_p53_BLA_p3_viability	Cell proliferation	Human	Intestinal	HCT116
Tox21_p53_BLA_p4_viability	Cell proliferation	Human	Intestinal	HCT116
Tox21_p53_BLA_p5_viability	Cell proliferation	Human	Intestinal	HCT116
Tox21_VDR_BLA_agonist_viability	Cell proliferation	Human	Kidney	HEK293T
Tox21_ESRE_BLA_viability	Cell proliferation	Human	Cervix	HeLa
Tox21_NFkB_BLA_agonist_viability	Cell proliferation	Human	Cervix	ME-180

$$\mathrm{SOF} = 1/(\mathrm{IP} - \mathrm{EA}) \quad (1)$$

$$\mathrm{EPH} = (\mathrm{IP} + \mathrm{EA})^2/8(\mathrm{IP} - \mathrm{EA}) \quad (2)$$

ROC AUC [20,33] was calculated based on the distributions of the "active" and "inactive" chemicals in the training set to advise variable selection. Fig. 1 shows that the six variables possess differentiated information in distinguishing between the "active" and "inactive" chemicals. The histogram for each design variable is provided in Fig. S3 in ESI.‡ To examine the dependency between the design variables, we computed the maximal information coefficients.[34,35] Notice (see Fig. 2) that IP, EPH and EA.aq contain higher degree of mutual information compared to the rest of the matrix elements. Simultaneously, the same three variables are ranked as the weakest predictors in the ROC analysis (Fig. 1). Thus, they form a less predictive group and were therefore excluded from further consideration.

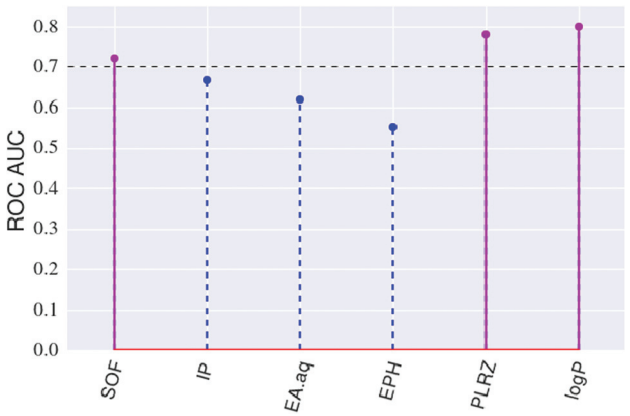

Fig. 1 ROC AUC between "active" and "inactive" chemicals in the training set. Selected design variables were inked in purple.

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

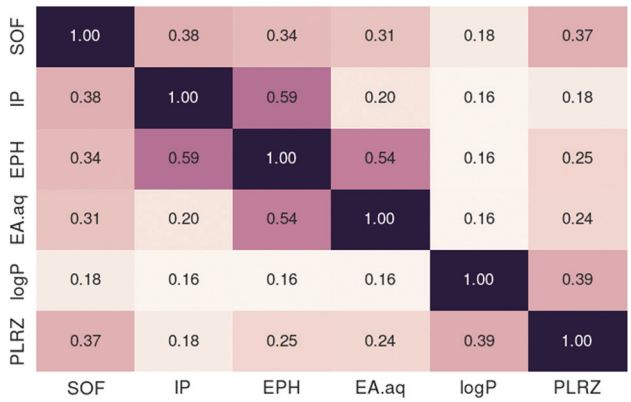

Fig. 2 Heatmap matrix for maximum information coefficients.

The remaining three design variables SOF, PLRZ and $\log P$ were retained for the Näive Bayesian model construction.

2.4. Predictive model construction and design guidelines extraction

The Näive Bayesian classifier[36] is an effective probabilistic classifier based on Bayes' theorem (eqn (3)) with independence assumptions between features.

$$\pi(\theta|X) = \frac{\pi(X|\theta)\pi(\theta)}{\int \pi(X|\theta)\pi(\theta)d\theta} \tag{3}$$

where θ denotes parameter, X denotes random variable, $\pi(\theta|X)$ denotes posterior probability, $\pi(\theta)$ denotes prior probability, $\pi(X|\theta)$ denotes the likelihood function, and the denominator integral denotes the marginal likelihood.

In this study, the parameter θ represents the class identifier and X represents the design variables obtained from the previous section. Our interest is to calculate the posterior or benign probability $\pi(\theta|X)$, the probability for a chemical to be "inactive". The challenge resides in estimating the likelihood function without the complete knowledge of the interactions between the X_i. The independence assumption allows one to express the likelihood function as a product of the conditional probabilities for each individual design variable. To examine the appropriateness of employing this assumption, mutual information (eqn (4)) between design variables was computed as shown in Fig. 2. The three selected variables (PLRZ, SOF and $\log P$) showed marginal dependency between each other. Therefore, the independent assumption can be reasonably adopted.

$$I(X,Y) = \int\int p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\mathrm{d}x\mathrm{d}y \tag{4}$$

where $p(x,y)$ is the joint probability density function of X and Y, and $p(x)$ and $p(y)$ are the marginal probabilities. Implementing the independence assumption also makes the probability function conveniently retrievable by solving its inverse function. In other words, for a given posterior probability, it is possible to go back to find corresponding solutions in the design variable space. The complete solution set is valuable in the design of safer, green molecules.

The model construction, data analysis and the graphical visualization in this study were coded using the Python programming language,[37,38] libraries and packages.[39–46]

3. Results and discussion

3.1. Mechanistic rationale for design variables

Creating and choosing appropriate design variables that encode the principles governing the mechanisms that lead to specific toxicity endpoints is the essence of green molecular design. It requires a chemo-physical theory to construct a design variable space that can effectively reflect the molecular initiation of toxicological events. The molecular basis for chemicals to incur toxicity is through intermolecular interactions between chemicals and critical biological targets.[47] For instance, covalent modification to proteins, especially thiol groups, has been recognized as a trigger for cellular toxicity.[48,49] In some instances chemical agents need to pass through cell membranes to reach their biological targets. In other cases, chemicals may cause cell lethality by destabilizing cell membranes themselves. Electrostatic interactions between chemicals and cellular membranes often contribute to this type of effect.[50]

Given potential molecular mechanisms through which chemicals can invoke toxicity, the next step is to collect design variables that properly describe or control those mechanisms. For instance, the hydrophobicity of a molecule is usually an important indicator of cell membrane permeability. $\log P$ is frequently used as a simple descriptor for molecular hydrophobicity in quantitative structure–activity relationship models (QSARs) to predict acute aquatic toxicity.[51] Cell membrane permeability can have two consequences. The direct one is that disruption of the membranes themselves is toxic, leading to necrosis. Alternatively, a chemical can pass through the cell membrane to enter cellular interior or interact with receptors on the membrane surface to exert a wide range of effects. These can be specific, receptor-mediated effects, which alter cell signaling, or non-specific effects which are more likely to lead to cell stress and cytotoxicity. The physical nature of these non-specific interactions is often mediated by induced electric dipole moments and dispersion forces. Polarizability is a physical quantity that describes the relative tendency of molecular electron cloud distortion under the influence of an external electric field. It quantifies the energy alteration upon molecular attractions.[52] Therefore, it is reasonable to include polarizability in modeling cytotoxicity. Hard–soft acid–base (HSAB) theory estimates the tendency for chemicals to form covalent bonds, which is linked to the toxic potential of certain chemicals.[53] Molecular softness built upon DFT provides a means to quantify this tendency.[54] Thus, molecular softness was included in the design variable space in this study.

 This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

Note that the exact primary cause of cytotoxicity is difficult to discover. We have thoroughly analyzed the correspondence between specific molecular interactions, generalized cell-stress and cytotoxicity in this data set[26] with multiple assays for each of multiple cell-stress processes (mitochondrial disruption, oxidative stress, ER stress, heat shock, apoptosis, *etc.*). What we almost universally see is that multiple cell stress assays are activated at roughly the same concentration (within the uncertainty of the assays themselves). We rarely ever see cytotoxicity in the absence of these more general cell stress markers, and only in the presence of some particular target activity (*e.g.* a receptor or enzyme). So we find (in most cases) that when cytotoxicity happens (in time and concentration), many or most of these other processes are also occurring, meaning that it is difficult to pull apart the specific cause–effect relationships between different cell-stress processes and cytotoxicity.

3.2. Performance evaluation for the predictive model

We used the following measures to evaluate the performance of the predictive model.

$$\text{Precision} = \text{true positive}/(\text{true positive} + \text{false positive}) \quad (5)$$

$$\text{Recall} = \text{true positive}/(\text{true positive} + \text{false negative}) \quad (6)$$

$$\text{F1 score} = 2(\text{precision} \times \text{recall})/(\text{precision} + \text{recall}) \quad (7)$$

We define "inactive" chemicals as positive in this study because they represent the group of interest from the perspective of green chemical design. Therefore, type I error deserves the best attention to be avoided. Precision signifies the success rate in identifying true positives out of all predicted positives. It is a very important criterion in model evaluation according to the precautionary principles – the probability that chemical is truly safe given that it is predicted to be safe should be maximized. Recall assesses the ability of a model to make positive predictions. Both of these variables depend on the positive ratio in the data set. F1 score as the harmonic mean between precision and recall represents a balanced assessment to the model performance. ROC AUC measures the overall ability for a model to separate "inactive" chemicals from "active" ones.

The Näive Bayesian model was trained on ~500 chemicals with 10-fold cross validation and externally tested on the remaining ~500 chemicals. The model performance is summarized in Table 2.

It is easy to notice from Table 2 that the three performance indicators are in reasonable agreement. Also, the cross-validation and external evaluation results agree well with each other, indicating a lack of overfitting. While it would be surprising to see overfitting with only three predictor variables in such a large and diverse chemical set, it is gratifying to achieve this level of predictivity. To best of the authors' knowledge, there are no other reports in the molecular design literature that have achieved this level of predictivity in the complete sample space. The high predictivity of this model indicates that a large majority of cytotoxicity is driven by relatively simple and non-specific molecular interactions rather than by a wide range of specific receptor-mediated mechanisms. It further suggests that the predictive model can be used for green design, at least to minimize the risk of cytotoxicity below 100 μM.

Table 2 Model performance evaluation

Measure	Precision	F1 score	ROC AUC
Cross validation	0.77	0.77	0.82
External evaluation	0.78	0.77	0.81

3.3. Probabilistic design diagram

Because of the Näive Bayes assumption, the inverse function of the Bayesian model (eqn (3)) can be derived with relative ease. To present the undetermined format of the solutions, we proposed a probabilistic diagram built upon a parametric description of function solution space.[55]

There are three advantages to using this probabilistic diagram to guide safer molecular design. Firstly, unlike the recursive partition strategy[14,15] where only partial solutions were provided, this diagram-based approach reveals the complete solution in the design variable space. This feature does not only boost the elegance of the model but also broadens its utility. For instance, chemists may be subject to constraints on certain design variables due to specified functional requirements. In those scenarios, a complete solution renders a higher flexibility for the users to seek solutions in the less restricted regions in the design variable space. Secondly, instead of setting a simple yes/no criterion, this diagram estimates probabilities for a chemical to be "inactive" (not cytotoxic in this study). This is considered to be a more realistic approximation of complex systems such as biology. Probabilities can make the chemists aware of the change in cytotoxicity potential while adjusting design variable values. Thus, it provides quantitative information in the decision making process. Thirdly, the diagram can be used in two directions, estimating benign probability or seeking solutions in the design variable space. As shown in Fig. 3, a sample solution is indicated by the dotted lines. In this case, one wants to design a chemical with about 82% probability of not invoking cytotoxicity. For certain functional reasons, the polarizability has been set near 12 cm^3 mol^{-1}. Given those conditions, the designer connects two points on the probability and PLRZ axes to arrive at a specific point on R_1 axis. Now one sees multiple options for the combination of $\log P$ and SOF to satisfy the two constraints above. The example on the graph illustrates one option: 2.8 for $\log P$ and 0.105 eV for SOF. The resultant molecule corresponds to propylparaben, a naturally occurring chemical that has been approved by US Food and Drug Administration (FDA) for safe use in cosmetics. Other possible solutions may exist and all can be found on the design diagram. In a word, the design diagram presents all the possible solutions in the design variable space for any particular chosen benign probability. Alternatively, it is possible to walk from the opposite direction on the graph. One can pre-assign values to all of the design variable axes and connect them to reach a resultant

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

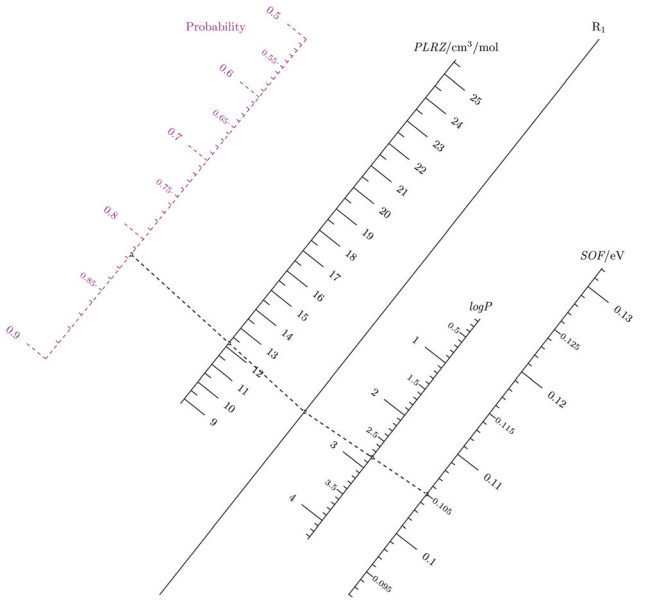

Fig. 3 Probabilistic design diagram for chemicals with customizable reduction of cytotoxicity potentials. The benign probability axis is printed purple while other design variables axes are printed black. R_1 indicates an auxiliary axis. The numerical range of each design variable axis are calculated using 5 to 95 percentiles of numerical data to produce probabilities corresponding to the range of the benign probability axis. Benign probability above 50% is considered for the green molecular design purpose.

benign probability. This feature allows one to assess the benign probability for an existing or new chemical with defined design variable values.

4. Conclusions

This study outlines a probabilistic strategy for safer chemical design. Cytotoxicity was chosen as a demonstration endpoint. By building design variables using physically meaningful and chemically intuitive attributes, employing the Näive Bayesian algorithm which allows for easy function inversion and presenting the full solution in the design space using a probabilistic diagram, this research addressed the three main challenges in molecular design at once. The resultant probabilistic diagram can guide the design of chemicals with a customized probability to reduce the risk of incurring cytotoxicity. This approach renders high level flexibility to chemists when seeking solutions in the design variable space. The probabilistic scales are useful in assessing the quantitative impact on benign probability while altering the numerical values of the design variables in practice. It is an expansion of the existing molecular design methods and will serve as a ground work for future green design research.

Acknowledgements

This research is supported by EPA/NSF Networks for Sustainable Molecular Design and Synthesis. PTA would like to express appreciation to QAFCO for funding support. The authors would like to thank for the helpful discussions with Dr Yan Zhang, Dr Imran Shah, Dr Declan Clarke, and Dr Philip Coish. The authors acknowledged the computational support provided by Dr William Jorgensen, Dr Julian Tirado-Rives and Yale high performance computing platform.

References

1 National Research Council (US) Steering Committee on Identification of Toxic and Potentially Toxic Chemicals for Consideration by the National Toxicology Program, *Toxicity Testing: Strategies to Determine Needs and Priorities*, National Academies Press, Washington DC, USA, 1984.
2 T. Hartung, *Nature*, 2009, **460**, 208–212.
3 Committee on Toxicity Testing and Assessment of Environmental Agents, Board on Environmental Studies and Toxicology, Institute for Laboratory Animal Research, Division on Earth and Life Studies, National Research Council, *Toxicity Testing in the 21st Century: A Vision and a Strategy*, National Academies Press, Washington DC, USA, 2007.
4 D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer and R. J. Kavlock, *Toxicol. Sci.*, 2006, **95**, 5–12.
5 F. S. Collins, G. M. Gray and J. R. Bucher, *Science*, 2008, **319**, 906–907.
6 R. S. Judson, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah, A. M. Richard and D. J. Dix, *Environ. Health Perspect.*, 2009, **118**, 485–492.
7 S. J. Shukla, R. Huang, C. P. Austin and M. Xia, *Drug Discovery Today*, 2010, **15**, 997–1007.
8 R. Kavlock, K. Chandler, K. Houck, S. Hunter, R. Judson, N. Kleinstreuer, T. Knudsen, M. Martin, S. Padilla, D. Reif, A. Richard, D. Rotroff, N. Sipes and D. Dix, *Chem. Res. Toxicol.*, 2012, **25**, 1287–1302.
9 M. S. Attene-Ramos, N. Miller, R. Huang, S. Michael, M. Itkin, R. J. Kavlock, C. P. Austin, P. Shinn, A. Simeonov, R. R. Tice and M. Xia, *Drug Discovery Today*, 2013, **18**, 716–723.
10 T. M. Martin, C. M. Grulke, D. M. Young, C. L. Russom, N. Y. Wang, C. R. Jackson and M. G. Barron, *J. Chem. Inf. Model.*, 2013, **53**, 2229–2239.
11 S. Ekins, *J. Pharmacol. Toxicol. Methods*, 2014, **69**, 115–140.
12 R. J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R. W. Setzer, I. Shah, D. Villeneuve and E. Weber, *Toxicol. Sci.*, 2008, **103**, 14–27.
13 P. T. Anastas and J. C. Warner, *Green Chemistry: Theory and Practice*, Oxford University Press, New York, USA, 1998.
14 A. M. Voutchkova, J. Kostal, J. B. Steinfeld, J. W. Emerson, B. W. Brooks, P. Anastas and J. B. Zimmerman, *Green Chem.*, 2011, **13**, 2373.
15 J. Kostal, A. Voutchkova-Kostal, P. T. Anastas and J. B. Zimmerman, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 6289–6294.
16 J. Kostal, A. Voutchkova-Kostal, B. Weeks, J. B. Zimmerman and P. T. Anastas, *Chem. Res. Toxicol.*, 2012, **25**, 2780–2787.

 This journal is © The Royal Society of Chemistry 2016

Open Access Article. Published on 05 July 2016. Downloaded on 1/19/2026 6:25:52 PM.
This article is licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported Licence.

17 J. A. Kramer, J. E. Sagartz and D. L. Morris, *Nat. Rev. Drug Discovery*, 2007, **6**, 636–649.
18 A. Schrage, K. Hempel, M. Schulz, S. N. Kolle, B. van Ravenzwaay and R. Landsiedel, *ATLA, Altern. Lab. Anim.*, 2011, **39**, 273.
19 N. C. Kleinstreuer, J. Yang, E. L. Berg, T. B. Knudsen, A. M. Richard, M. T. Martin, D. M. Reif, R. S. Judson, M. Polokoff, D. J. Dix, R. J. Kavlock and K. A. Houck, *Nat. Biotechnol.*, 2014, **32**, 583–591.
20 T. A. Lasko, J. G. Bhagwat, K. H. Zou and L. Ohno-Machado, *J. Biomed. Inf.*, 2005, **38**, 404–415.
21 S. R. Langdon, J. Mulgrew, G. V. Paolini and W. P. Van Hoorn, *J. Cheminf.*, 2010, **2**, 11.
22 A. C. Lee, K. Shedden, G. R. Rosania and G. M. Crippen, *J. Chem. Inf. Model.*, 2008, **48**, 1379–1388.
23 L. Molnár, G. M. Keseru, A. Papp, Z. Lorincz, G. Ambrus and F. Darvas, *Bioorg. Med. Chem. Lett.*, 2006, **16**, 1037–1039.
24 M.-L. Zhao, J.-J. Yin, M.-L. Li, Y. Xue and Y. Guo, *Interdiscip. Sci.*, 2011, **3**, 121–127.
25 U. EPA, *ToxCast & Tox21 Summary Files Released Dec. 2014.*, http://www.epa.gov/chemical-research/toxicity-forecaster-toxcasttm-data.
26 R. Judson, K. Houck, M. Martin, A. M. Richard, B. Thomas, I. Shah, S. Little, J. Wambaugh, R. Woodrow, P. Kothya, J. Phuong, D. Filer, D. Smith, D. Rotroff, N. Kleinstreuer, N. Sipes and M. Xia, *Toxicol. Sci.*, 2016, DOI: 10.1093/toxsci/kfw092.
27 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.
28 *Marvin Calculator Plugins*, Calculator Plugins were used for structure property prediction and calculation, Marvin 6.3.4, 2013, ChemAxon (http://www.chemaxon.com).
29 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski and D. J. Fox, *Gaussian ~09 Revision D.01*, Gaussian Inc., Wallingford CT, 2009.
30 A. D. Boese and J. M. L. Martin, *J. Chem. Phys.*, 2004, **121**, 3405–3416.
31 A. V. Marenich, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. B*, 2009, **113**, 6378–6396.
32 R. G. Parr, L. V. Szentpály and S. Liu, *J. Am. Chem. Soc.*, 1999, **121**, 1922–1924.
33 A. P. Bradley, *Pattern Recognit.*, 1997, **30**, 1145–1159.
34 D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher and P. C. Sabeti, *Science*, 2011, **334**, 1518–1524.
35 D. Albanese, M. Filosi, R. Visintainer, S. Riccadonna, G. Jurman and C. Furlanello, *Bioinformatics*, 2013, **29**, 407–408.
36 G. H. John and P. Langley, Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 1995, pp. 338–345.
37 T. E. Oliphant, *Comput. Sci. Eng.*, 2007, **9**, 10–20.
38 K. J. Millman and M. Aivazis, *Comput. Sci. Eng.*, 2011, **13**, 9–12.
39 E. Jones, T. Oliphant, P. Peterson, *et al.*, *SciPy: Open source scientific tools for Python*, 2001, http://www.scipy.org/, [online; accessed 2016-01-24].
40 S. van der Walt, S. Colbert and G. Varoquaux, *Comput. Sci. Eng.*, 2011, **13**, 22–30.
41 *pandas: Python Data Analysis Library*, online, 2012, http://pandas.pydata.org/.
42 F. Pérez and B. E. Granger, *Comput. Sci. Eng.*, 2007, **9**, 21–29.
43 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.
44 J. D. Hunter, *Comput. Sci. Eng.*, 2007, **9**, 90–95.
45 M. Waskom, O. Botvinnik, P. Hobson, J. Warmenhoven, J. B. Cole, Y. Halchenko, J. Vanderplas, S. Hoyer, S. Villalba, E. Quintero, A. Miles, T. Augspurger, T. Yarkoni, C. Evans, D. Wehner, L. Rocher, T. Megies, L. P. Coelho, E. Ziegler, T. Hoppe, S. Seabold, S. Pascual, P. Cloud, M. Koskinen, C. Hausler, Kjemmett, D. Milajevs, A. Qalieh, D. Allan and K. Meyer, *seaborn: v0.6.0*, 2015, DOI: 10.5281/zenodo.19108.
46 L. Roschier, *PyNomo - a program to create nomographs with Python, 2009*, http://sourceforge.net/projects/pynomo/.
47 D. V. Parke, *Regul. Toxicol. Pharmacol.*, 1982, **2**, 267–286.
48 J. A. Hinson and D. W. Roberts, *Annu. Rev. Pharmacol. Toxicol.*, 1992, **32**, 471–510.
49 R. M. LoPachin and A. P. DeCaprio, *Toxicol. Sci.*, 2005, **86**, 214–225.
50 H. C. Shertzer, M. Sainsbury, P. R. Graupner and M. L. Berger, *Chem.-Biol. Interact.*, 1991, **78**, 123–141.
51 M. D. Cronin, *Curr. Comput.-Aided Drug Des.*, 2006, **2**, 405–413.
52 D. M. Quinn, H. K. Nair, J. Seravalli, K. Lee, A. T. Arbuckle, Z. Radić, V. D. C. Vellom, P. N. Pickering and P. Taylor, in *London Dispersion Interactions in Molecular Recognition by Acetylcholinesterase*, ed. D. M. Quinn, A. S. Balasubramanian, B. P. Doctor and P. Taylor, Springer, 1995.
53 R. M. LoPachin and T. Gavin, *Chem. Res. Toxicol.*, 2014, **27**, 1081–1091.
54 R. G. Parr and R. G. Pearson, *J. Am. Chem. Soc.*, 1983, **105**, 7512–7516.
55 L. I. Epstein, *Nomography*, Interscience Publishers, INC., New York, USA, 1958.