nature communications

Perspective

https://doi.org/10.1038/s41467-023-40459-8

Applied machine learning as a driver for polymeric biomaterials design

Received: 8 February 2023

Accepted: 24 July 2023

Published online: 10 August 2023

Check for updates

Samantha M. McDonald[1], Emily K. Augustine[2], Quinn Lanners[3], Cynthia Rudin[3], L. Catherine Brinson[2] & Matthew L. Becker[1,2]

Polymers are ubiquitous to almost every aspect of modern society and their use in medical products is similarly pervasive. Despite this, the diversity in commercial polymers used in medicine is stunningly low. Considerable time and resources have been extended over the years towards the development of new polymeric biomaterials which address unmet needs left by the current generation of medical-grade polymers. Machine learning (ML) presents an unprecedented opportunity in this field to bypass the need for trial-and-error synthesis, thus reducing the time and resources invested into new discoveries critical for advancing medical treatments. Current efforts pioneering applied ML in polymer design have employed combinatorial and high throughput experimental design to address data availability concerns. However, the lack of available and standardized characterization of parameters relevant to medicine, including degradation time and biocompatibility, represents a nearly insurmountable obstacle to ML-aided design of biomaterials. Herein, we identify a gap at the intersection of applied ML and biomedical polymer design, highlight current works at this junction more broadly and provide an outlook on challenges and future directions.

Many of the machine learning (ML) approaches at the intersection of medicine and chemistry focus on small molecule synthesis for drug discovery[1–7]. As shown in Fig. 1, there is a considerable gap in strategies targeting polymers in medicine despite medical polymers representing an 18.4-billion-US dollar global market as of 2021– appearing in diverse applications such as catheters, coatings, implants, etc[8–10].

Among the bottlenecks experienced in the development of commercial medical polymers, the research and design of such materials represent a huge investment of time, money, and energy. New materials are often designed through experimental intuition and further developed through trial-and-error synthesis. This process is not only economically inefficient, but also often fails to produce polymers that have all the target properties for an intended application. This is due, in part, to the fact that polymers exhibit less predictable/intuitive structure-property relationships than small molecule counterparts due to the greater number of variables which dictate their properties—namely, composition, molecular mass, intermolecular forces, and architecture[11,12]. Even within a class of compositionally similar polymers, these relationships can be nonlinear and hard to identify, which makes it difficult to design materials for a specific property outcome. Thus, employing ML as a tool that is sensitive to patterns in data which are critically important, but indiscernible to humans, could accelerate the development of translationally relevant polymers[13].

Most work at the intersection of ML and polymer design more broadly can be separated into two general tasks:

1. *Property prediction (or forward problem design):* Given a polymer structure, predict specified properties of interest. This approach is highly valuable for the screening of candidate materials, enabling synthesis of the most promising candidate rather than the whole library.
2. *Structure generation (or inverse problem design):* Given properties of interest, predict polymeric structures that may demonstrate

[1]Department of Chemistry, Duke University, Durham, NC, USA. [2]Thomas Lord Department of Mechanical Engineering and Materials Science, Duke University, Durham, NC, USA. [3]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA. e-mail: matthew.l.becker@duke.edu

https://doi.org/10.1038/s41467-023-40459-8

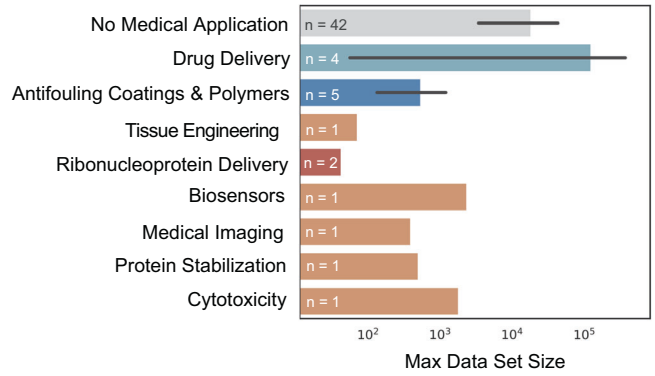

Fig. 1 | Data set sizes documented for applications within polymer chemistry. There are only a few examples of ML applied to biomedical polymer questions (n = the number of papers for each application). Data availability is a considerable concern for ML approaches within many biomedical applications as demonstrated by the relatively few number of papers which have been published on this topic as well as the small size of the data sets used in these existing approaches.

the desired properties. This strategy can expedite the materials design process and remove limitations set by user intuition.

Applications of ML are not limited to polymer design and have also been successfully implemented for process optimization and advancing our understanding of structure-property relationships within these complex systems. Ultimately, accessing the next generation of polymeric biomaterials will more than likely require some combination of these approaches. To quote Anne Fischer, a DARPA program manager for accelerated molecular discovery:

"It's not about replacing chemists. It's about giving chemists the tools to allow them to implement and apply the chemistry and allow them to be creative high-level thinkers."[14]

Challenges with data availability

Advancements in property prediction and inverse modeling in related fields have been driven primarily by relatively data-hungry supervised learning algorithms, including deep learning, random forests, gaussian process models[15–18]. Such models are very promising for developing advancements in polymer design given the complexity of polymer systems. However, they often require more labeled data than the typical experimentalist produces to discover more fine-tuned interactions with limited prior understanding of the system.

As such, data availability represents a primary obstacle for this field irrespective of the approach[13]. Experimental datasets are often small (on the order of 1–20 unique structures) and incompatible with each other due to differences in experimental methods and/or data analysis. Property handbooks which serve as reliable references for the experimentalist are often compiled with only the polymer name or with limited structural information[19,20]. Polymers are not always named according to the conventions of the International Union of Pure and Applied Chemistry, which makes the name a poor representation of the materials' structure. Additionally, these sources typically are not easily exportable, which is necessary for implementation in code. Online databases, as shown in Table 1, contain thousands of polymers with some of their associated properties, structural images, and some additional methods of identification[13,21–23]. However, both polymer handbooks and online databases suffer from high data sparsity and are limited in the properties they contain. This limits the properties that can be predicted with supervised learning. For example, it would be difficult to create a labeled dataset for training a model to predict degradation time given that this property is often not included in current databases. This also affects what properties can be used as input features (e.g., molecular mass) which may limit prediction

Table 1 | Summary of current online databases

Database[Ref]	# of Polymers	Polymer Class(es)	Properties	Drawbacks
PolyInfo[113a]	31,495	Diverse	Physical, optical, thermal, electrical, physicochemical, rheological, solution, mechanical	●, ✤
Khazana[23b]	965	Conjugated polymers, commercial thermoplastics, polyesters	Electrical, optical	★
Polymers: a Property Database[114a]	30,000	Diverse	Physical, optical, thermal, electrical, physicochemical, rheological, mechanical	●, ★, ◀
Polymer Property Predictor and Database[115a]	212–6524	Conjugated polymers, commercial thermoplastics	Solution, thermal	✤
MatWeb[116a]	97,635	Commercially available polymers	Physical, optical, thermal, electrical, physicochemical, rheological, mechanical	●, ★
Block Copolymer Phase Behavior Database[117a]	5300	Block copolymers	Phase measurements for block copolymers	-
Electron Affinity and Ionization Potential Data[62c]	42,966	Halogenated polymers, conjugated polymers, charged polymers	Copolymer electron affinities and ionization potentials	-

● = fee or academic affiliation required, ✤ = no public API or ability to download, ◆ = no structure information, ◀ = unstandardized/text entries, ★ = cannot download the whole dataset/can only download a subset.
[a]experimental data.
[b]experimental and simulated data.
[c]simulated data.

performance. Additionally, these data sources are often not accessible −frequently requiring an academic affiliation, have no easy way to download the data, and/or expect a fee before use.

Experimental datasets containing properties of interest (e.g., in vivo degradation time, cytotoxicity) for the design of biomedical polymers specifically demonstrate high scarcity with respect to the number and size of available datasets. Figure 1 shows efforts in ML applied to biomedical polymers by application. This scarcity persists for many reasons, including the time and cost associated with characterizing in vivo properties, lack of standardization within in vitro methods, and the lack of applicability of some biomedical properties in other fields of interest to polymers (e.g., drug release profile has little applicability for energy or sustainable materials).

Previous work in ML applied to polymer design has addressed data availability concerns by simulating data[24–26]. This strategy is particularly promising for generating labeled datasets for properties which are not commonly characterized because they are difficult to characterize or only relevant for niche applications. Batra et al. used simulated data to train a property prediction algorithm as part of an inverse design approach. In particular, they used density functional theory composition to simulate polymer bandgap labels, one of the target properties in their approach, since experimental values of this property are less common in the literature[25]. Existing molecular dynamics simulations of biomedical properties, such as cytotoxicity, could be used to generate datasets for supervised ML and demonstrate the utility of simulating other medical properties[27,28]. However, it should be noted that this approach can result in error propagation through to the final algorithm, so special care should be taken to confirm the fidelity of the simulation results.

Transfer learning builds a ML model by using a smaller dataset to finetune or adapt a model that was originally trained for a similar task on a larger dataset[29]. This approach has been extremely successful in other fields and within chemistry for overcoming data limitations and improving model performance[29–32]. Using transfer learning, a model pretrained on a large simulated dataset can be finetuned using a smaller experimental dataset to address error propagation concerns, while still reducing the amount of necessary experimental data. Examples of transfer learning for scarce properties from models trained on physically related properties also present a route to developing stronger models with less experimental data[33]. However, transfer learning still often requires more data than the typical experimentalist generates.

Other approaches have leveraged high throughput or automated experiments to generate data more quickly than traditional synthesis[34–36]. The most common high-throughput strategies use continuous-flow systems, plate-based methods, or reactor arrays to run many polymerization reactions simultaneously[37–44]. However, other experimental setups, such as microfluidic reactors or PCR thermocycler setups, may be favorable depending on the property of interest and the number of variables which must be controlled[45,46]. These principles can be automated via programmable robots to further eliminate the need for human intervention and to run reactions at times unfavorable for researchers[43,44,47]. High-throughput approaches are generally promising for properties which can be measured directly with small quantities of polymer (e.g., water uptake, cytotoxicity) and can be combined with methods which can accommodate relatively small datasets, such as active learning or Bayesian optimization[36,48,49]. Within regenerative medicine, high-throughput methods have shown promise for the efficient generation of polymer excipients, and antimicrobial polymer discovery[40–42]. Generating datasets via these existing approaches could be a more immediately achievable way to apply ML to biomedical polymer design. However, these strategies may be hard to scale for properties which require larger quantities of polymer and/or which require processing steps, but can still be employed to reduce the active synthesis time for experimentalists.

Additionally, chain-growth polymerizations, such as ring-opening polymerization and reversible addition fragmentation chain transfer polymerization, have been the methods of choice for high-throughput experiments[37–46]. There is a notable gap of high-throughput strategies for step-growth polymerizations even though these structures are common among commercially available medical polymers and are of interest for developing new medical polymers due to backbone heteroatoms which enable degradability. Poly(lactic acid) (PLA), poly(-urethanes) and nylons are classes of polymers commonly synthesized via step-growth polymerization methods and employed in a diverse range of commercial medical applications, including medical tubing, short-term implants, and sutures[50–53]. Despite this, there are only a couple of examples of high-throughput strategies which target step-growth type reactions[47,54]. Thus, the development of scalable high-throughput methods which include step-growth polymerizations and biomedical properties would greatly advance the available data for applied ML.

Ultimately, the need for an accessible, high-quality data source is undeniable. The recently released Community Resource for Innovation in Polymer Technology (CRIPT) is one potentially scalable architecture which shows immense promise toward this goal, focusing on the curation of current and future data rather than trying to solve the curation of historical data[55]. However, the success of this approach is contingent on widespread contribution to the database by experimentalists and the current number of polymers contained in the database is unclear. Additionally, emerging databases must still address polymer naming inconsistencies which has driven the use of BigSMILES strings and knowledge graphs as labels in lieu of traditional names[56]. The adoption of alternative labeling conventions should be widespread to facilitate interoperability between databases and researchers' familiarity with these new labels. Guidelines for and strong examples of data-sharing are outlined in the FAIR principles. Briefly, these principles identify good data-sharing platforms as being: Findable, Accessible, Interoperable, and Reusable[57]. Updating existing platforms to abide by these principles would considerably lower the barrier to applying ML to a broader set of polymer design problems. Similarly, these principles should be at the core of future data-sharing ventures.

Encoding chemical information

Encoding the chemical information of polymers into a machine-readable format is an essential and nontrivial step for ML in this area. Polymer structures are complex inputs which have been represented through molecular graphs, monomer simplified molecular-input line-entry system (SMILES), and polymer SMILES representations (BigSMILES, curlySMILES, etc)[13,58,59]. SMILES representations have been a popular choice as they can be one-hot encoded (a group of bits among which the combinations of values are only those with a single high [1] bit and all the others low [0]) and are more easily interpreted by experimentalists who may not have strong computer science skills. Most approaches opt for training algorithms on repeat unit SMILES which makes it difficult to deal with structures which contain more than one repeat unit (e.g., copolymers) or exhibit complex architecture (e.g., star-shape)[33,60]. BigSMILES and curlySMILES represent polymer extensions which factor in the limitations intrinsic to the original SMILES syntax. However, these representations are not all-inclusive and may require passing additional variables, such as monomer stoichiometry and polymer dispersity. Similarly, the encoding of SMILES strings results in a high dimensional input which is not suitable for all ML approaches depending on the complexity of the model and data set size.

Molecular descriptors can be used in lieu of an encoded SMILES input to reduce the number of input features. This approach works well for systems which have low structural diversity and/or with algorithms that cannot accommodate encodings with connectivity

Table 2 | Python polymer packages

Entry	Package[Ref]	Description
E1	XenonPy[32]	Various pretrained models for different properties of interest and other ML tools related to chemical problems (including polymers)
E2	Topoly[118], PSP[119], PysoftK[120]	Toolkits for polymer topology and structure
E3	pysimm[121], Polyply[122], m2p[123]	Open-source polymer simulations and chain generation
E4	PMD[124]	High throughput molecular dynamic simulations for a variety of properties
E5	BigSMILES Parser[125]	Parses BigSMILES strings
E6	Matminer[126–138]	Materials data miner from online databases

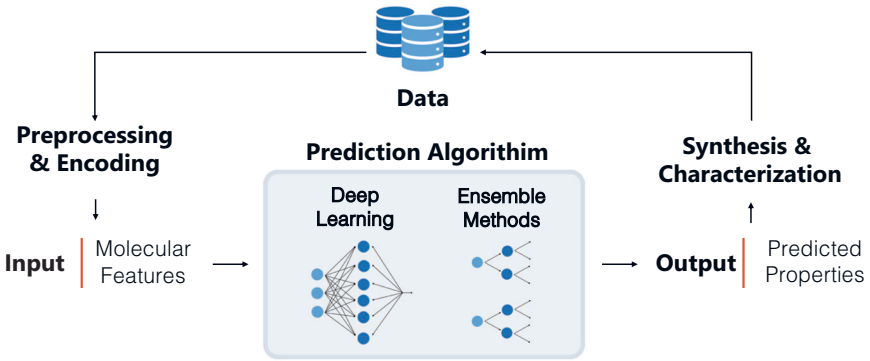

Fig. 2 | General ML workflow for property prediction tasks. Data (i.e., polymers with known properties) must be preprocessed and encoded before passing desired input (e.g., encoded chemical structure, molecular descriptors) into a prediction algorithm. Irrespective of algorithm choice, training proceeds by tuning the model hyperparameters to minimize prediction error. The trained algorithm can then be used to screen polymer candidates prior to experimental synthesis & characterization. While deep learning and ensemble methods are the most widely used, other supervised methods have been employed (see Table 3) and may be preferred based on the application.

information, such as linear models, support vector machines and random forest[61]. As discussed in the *Potentially Helpful Tools* section, existing Python packages have been developed to automate the generation of these descriptors.

Graph representations have also demonstrated promising success as featurization methods which capture other polymer features (e.g., monomer stoichiometry, molar mass distribution) in addition to connectivity information[62]. While these methods are less interpretable than string counterparts, they are more complete representations of polymer systems.

Potentially helpful tools

An abundance of Python packages for chemistry have been comprehensively curated into a repository on GitHub[63]. Many tasks have been tackled through these packages including, but not limited to structure representation (e.g., RDkit), database wrappers (e.g., pubchempy, ChemSpiPy) and atomistic simulations. Table 2 shows a collection of Python packages focused particularly on polymer chemistry. E3 and E4, may be helpful in generating simulated datasets. Other packages like XenonPy or the BigSMILES parser (Table 2 E1 and E5) can be incorporated into approaches directly to reduce development time.

Current work in ML applied to polymer chemistry

Property prediction and inverse design, as discussed in the following section, represent the two most prominent supervised tasks within machine learning applied to polymer chemistry. Unsupervised methods, such as self-organizing maps, are much less common within polymer chemistry but can be used to explore structure-property correlations and visualize high dimensional data[36,64]. While this section does not discuss implementation at length, Meyer et al. present a thorough tutorial on how to apply ML to questions within polymeric biomaterials and model selection has been discussed more broadly elsewhere[65–67]. While Python is the language of choice for ML applied to chemical questions, other languages, like R and C++, can be used to build ML models. It should also be noted that ML methods are in an era of explosive growth, which may provide creative solutions to problems not yet solved within polymer chemistry.

Property prediction

Property prediction tasks (workflow described in Fig. 2) represent the most explored area of this emerging field due to the straightforward problem statement, the accessibility of easy to implement algorithms, and the potential to achieve useful results even with smaller datasets (Table 3). These tasks can be considered on their own as proof-of-concept work, implemented for candidate screening and/or used to develop quantitative structure property relationships. It should be noted that it is important to be careful drawing quantitative structure-property relationships from predictive models because, unless carefully designed, model feature importance alone does not signify causal importance. As shown in Table 3, thermal properties are a popular target in property prediction tasks due to their relevance to a variety of fields, a greater degree of standardization between data sources, and their characterization for most new polymers in the literature. Outside of commonly characterized properties, property prediction tasks are only beginning to be applied to properties of exclusive interest to biomedically relevant polymers, physiological degradation time and biocompatibility (see ML for polymer chemistry in medical applications section).

Random forest and recurrent neural network (RNN) algorithms have been the primary choice for polymer property prediction tasks (see Table 3). This is symptomatic of the input data and the size of the chosen data set. Random forests and other ensemble methods can discover nonlinear interactions while often not requiring as much data as deep learning algorithms to perform well. Additionally, RNNs perform well with textual input data, which is advantageous for approaches which use a SMILES string to encode polymer structures. In both

Table 3 | Summary of approaches towards property prediction

Property Class	Properties	Data Set Size	Algorithms Used
Thermal	Glass Transition Temperature (T_g)[39,66,89-96]	43 – 17,001	RNN, LSTM, LR, DNN, PLS, SVM, RF, RecNN, e-SUSI, GPR
	Melt Transition Temperature (T_m)[39,90]	942 – 12,374	DNN, RF
	Specific Heat Capacity[39]	58 - 133,885	DNN, TL
	Thermal Conductivity[39]	332	DNN
Mechanical	Tensile Modulus[90]	306	LR, SVM, RF
	T_α/T_γ ratio[11]	440	RNN, MLP
	Dynamic Elastic Modulus[11]	440	RNN, MLP
Solution	Density[39,89,90]	48 – 8613	DNN, PLS, LR, SVM, RF
	Dissolution Parameter[89]	48	PLS
	Cloud Point[97]	171	GBDT
Electrical	Dielectric Constant[98]	1140	RNN
	Electron Affinity[29]	42,966	wD-MPNN
	Ionization Potential[29]	42,966	wD-MPNN
Intrinsic	Viscosity Average Molecular Mass[99]	118	LR, ANN
	Functional Group Indices[99]	81 – 111	LR, ANN
Optical	Refractive Index[100]	527	GPR
Barrier	Gas Permeability[101]	376 - 698	GPR

RNN = recurrent neural network, *LSTM* = long short-term memory , *LR* = linear regression, *DNN* = deep neural network, *PLS* = partial least squares regression, *SVM* = support vector machine, *RF* = random forest, *RecNN* = recursive neural network, *MLP* = multilayer perceptron, *e-SUSI* = ensemble supervised self-organizing mapping, *ANN* = artificial neural network, *GPR* = gaussian process regression, *GBDT* = gradient-boosted decision tree, *TL* = transfer learning, *wD-MPNN* = weighted directed message passing neural network, **linear method**, **ensemble method**, **deep learning**, **other methods**.

cases, existing Python packages facilitate the easy implementation of these models as much of the mathematical 'nitty gritty' has been abstracted away, allowing users to simply call and apply the algorithms to a dataset of choice. However, ease of implementation should be balanced with an understanding of the limitations associated with the chosen ML method to ensure correct usage.

Investigation into other ML approaches could expand this task with respect to model performance and/or the accessibility of the results. Physics-informed models can improve prediction accuracy and model robustness in smaller data regimes[68,69]. The underlying physics of a system can be introduced through data augmentation. Interventions in the model architecture can also be guided by domain knowledge, and/or introducing constraints (often on the model's loss function) based on what is physically possible[69]. For example, Bradford et al. included the Arrhenius equation in the final layer of their predictive model to improve the performance of ionic conductivity predictions for polymer electrolytes[68]. These types of improvements to the model architecture would benefit greatly from collaboration between polymer chemists and ML experts.

Compared to black-box approaches (i.e., models which do not provide information about feature importance like deep learning methods), interpretable ML algorithms (i.e., models which information about feature importance can be extracted like random forest algorithms) create models that are more transparent, easier to troubleshoot, and facilitate understanding of the chemical system[70,71]. Alternatively, black-box methods can be supplemented with models like Shapely Additive exPlanations (SHAP) and Locally Interpretable Model-agnostic Explanations (LIME) to assist in the interpretation of predictions[72,73]. These modeling approaches can identify potential structure-property relationships and elucidate complex design rules. However, experimentalists must note that predictive ML models cannot be used to make causal claims and any insight provided by such models should be validated through experimentation. Conversely, causal ML algorithms are designed to discover causal relationships[74]. These methods can be beneficial for causal inference but are often more difficult to implement and are only valid in certain settings. Thus, researchers should ensure that they fully understand how to properly implement these methods before using them for causal discovery. Within polymer design for regenerative medicine, Kumar et al. demonstrated the utility of SHAP and causal ML algorithms for identifying structure-property relationships associated with genetic cargo delivery from polymer systems[73]. Application of interpretable models, interpretation methods, and causal ML should be extended to other applications for biomedical polymers to further advance domain knowledge within these systems.

Structure generation

New polymer structures are often conceived from experimentalist intuition given their knowledge of existing work. Where property prediction tasks can be employed to screen candidates for accelerated materials design, the inverse design approach can train an algorithm which generates structures given desired properties (Fig. 3). This

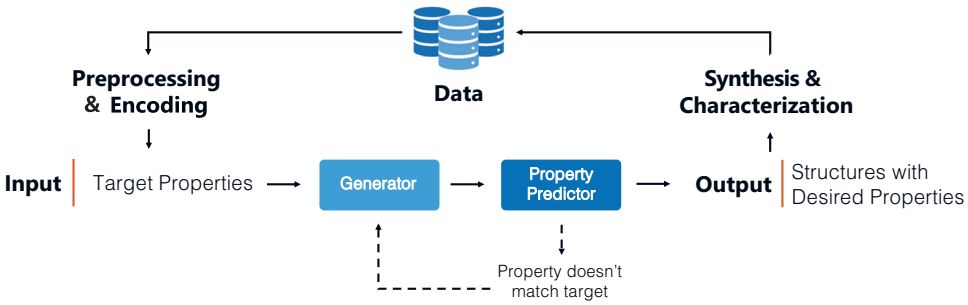

Fig. 3 | General ML workflow for inverse design approaches. Inverse design of polymers target algorithms which generate new, valid polymer structures with desired properties from property inputs. As seen in property prediction, data must be preprocessed & encoded prior to inverse design. Training involves the generation of a new structure through sequence perturbations or interpolations within existing latent spaces (represented here as 'generator'). The properties of the new structure are predicted and compared to the target properties (shown as 'property predictor'). The algorithms then iterate between these stages until structures with desired properties are achieved. While training of the 'generator' and 'property predictor' are approach-dependent, their hyperparameters may be tuned by minimizing the prediction error.

approach can overcome constraints introduced by the human imagination and preconceived beliefs about the complex structure-property relationships intrinsic to polymers. However, structure generation represents the most data hungry task due to the complexity of the models which can accomplish it; thus, big data availability presents a formidable obstacle.

The implementation of these inverse design approaches has been well discussed in other perspectives[75,76]. Briefly, the current efforts in the inverse design of polymer materials have followed a general workflow which iterates between generating new structures and predicting their properties (Fig. 3). This process proceeds by comparing the predicted properties to the target properties to minimize the difference between them. New structures can be generated through a sequence perturbation approach (e.g., Monte Carlo tree search, sequential Monte Carlo method, genetic algorithm) where a structure is incrementally changed to reach the property objective ("Generator" in Fig. 3)[33,77,78]. Alternatively, unsupervised methods (e.g., variational autoencoder) can be used to learn the structure-property latent space[25]. An interpolation method (e.g., linear interpolation) can then be used within regions which show higher probabilities for satisfying the property requirements to generate new structures. These candidate generation methods are then paired with other ML algorithms and/or molecular dynamics methods to predict their properties ("Property Predictor" in Fig. 3).

While success with these strategies has required considerable data (900–14,000 polymers), narrowing the scope of the inverse design problem to one class of polymers has proven successful on datasets as small as 171 polymers[25,33,78,79]. These current inverse design approaches have proceeded via similar workflows; thus, investigation into other inverse design algorithms (e.g., generative adversarial networks) represents a way to expand this field[80]. Ultimately, inverse design has the potential to accelerate the discovery of state-of-the-art materials, but is contingent on the creation of larger, more comprehensive datasets and/or the development of methods which can tolerate smaller, potentially imbalanced datasets.

ML for polymer chemistry in medical applications

Current work. Mathematical modeling has a well-established role in understanding the drug release profile from polymer systems; thus, it is no surprise that work at the intersection of ML and medically-relevant polymers has focused on drug release behaviors (Fig. 1)[81–85]. These approaches primarily target candidate screening methods through property prediction algorithms including perturbation theory machine learning (PTML), light gradient boosting machine (GBM), bagged multivariate adaptive regression splines (MARS), and random forests. Additionally, property prediction has been a dominant approach toward predicting surface adsorption/attachment behavior for antifouling coatings to aid in candidate screening or elucidating quantitative structure property relationships[86–88].

ML has also been applied toward the 3D printing conditions of tissue engineering scaffolds, 3D printing conditions of carbon doped polylactic acid for implantable biosensors and favorable conditions to generate medically-relevant microparticles[89,90]. Each of these approaches is underpinned by a focus on process optimization applied to a very narrow class of polymers with respect to structural diversity. Narrowing the scope of the input data accommodates a smaller number of observations which is compatible with the reality that biomedical properties are not as widely characterized as thermomechanical properties. Process optimization approaches also use less complicated inputs, such as concentration and printing speed, which facilitates the generation of larger datasets, more quickly when compared to chemical structure inputs which require synthesis to generate new observations.

Examples of ML applied to the synthesis of new biomedical polymers are scarce (Fig. 4). Efforts in high-throughput, combinatorial design of copolymers have led to discoveries in polymer-mediated ribonucleoprotein delivery, antibiofouling hydrogels, polymer-protein hybrids, and ^{19}F MRI agents[36,91–93]. These approaches leverage a limited compositional space and fast, simultaneous experiments to overcome data availability concerns. Additionally, as demonstrated by Reis et al., active learning can improve model performance considerably with only a few additional observations[36]. The compatibility of active learning with current experimental workflows makes it a promising way to introduce machine learning into existing polymer design problems.

Areas of need. The prediction of polymer properties that are unique to biomedical applications has lagged behind other achievements in ML property prediction (Fig. 4). Designing novel biocompatible polymers would benefit greatly from the ability to accurately predict cytotoxicity and bioactivity as demonstrated by previous work predicting nanoparticle toxicity[94,95]. The long timescale of data collection and non-standardized experimental methods to determine biocompatibility limit the generation of ML-viable datasets (Fig. 1 shows the relatively small size of biomedical data sets compared to non-biomedical applications).

Additionally, understanding the relationship between polymer degradation rate and underlying chemical structure is essential to tailoring biomedical polymers to their functionality and lifetime in the body. Physiologically degradable polymer systems improve patient quality of life by eliminating the need for secondary surgeries and reducing the possibility of long-term complications, such as infection

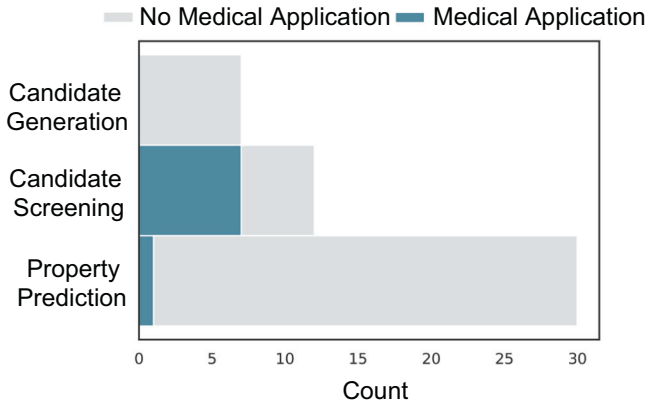

Fig. 4 | Uses for ML applied to polymer design. In each case, biomedical polymers represent a minority of the approaches and are notably missing from candidate generation strategies. The candidate screening methods shown here use property prediction algorithms to choose promising polymers out of a subset of interest. Thus, they differ from the property prediction category only in that their implementation goes a step further.

at the implant site. However, there are often differing degradation behaviors in vivo vs. in vitro which adds to the necessary time and resources required to develop new physiologically degradable polymers[96,97]. Inverse design approaches will greatly expediate the process of finding polymer structures with the desired biocompatibility and degradation timescale. Investigation into other ML methods which accommodate smaller datasets such as active learning, physics-informed and interpretable models represent current steps that can be taken to augment existing design approaches. Interpretable models, particularly, are promising for advancing our fundamental knowledge of the structure-property relationships which govern these systems in addition to augmenting experimental efficiency.

There is significant potential for the ML methods described herein to accelerate discoveries for biocompatible polymers in a variety of biomedical research areas.

Drug delivery and polymer excipients. Extensive research has focused on developing drug release systems that pair highly tunable polymeric carriers (e.g. hydrogels, films, fibers), with target small molecules[98–100]. The interaction of the carrier with the desired drug greatly impacts drug loading, long-term shelf stability, and ultimately release profile[101]. Polymers are also employed as excipients in drug formulations which require materials that can reduce drug-drug interactions[42,102]. Whereas current work has focused on adjusting parameters of existing polymer systems, little emphasis has been placed on ML polymer design with precise drug release profiles, degradation behavior, and stimuli-responsiveness. Drug delivery and polymer excipients represent a promising area for ML to flourish in polymer design due to the small polymer quantity requirements associated with characterizing the properties of interest. In the short-term, existing high-throughput methods for the design of polymer excipients by Mann et al. could be extended to generate a dataset large enough for ML to be applied meaningfully[42].

Regenerative medicine scaffolds. ML has been employed in the process optimization of scaffold preparation (e.g., bioprinting, electrospun fibers), but remains largely unexplored with regards to compositional landscape. Notably, there is a lack of biocompatible conducting materials for use in tissue scaffolds that mimic native tissue electrical properties[67]. It is also particularly challenging to quantify, and thus learn/predict some of the critical properties of polymeric scaffolds, including cellular proliferation and differentiation[103,104]. In particular, properties of interest to scaffolds for ML prediction will include cellular adhesion and in-growth, porosity, and tissue mechanics[105]. While data availability limits what can currently be done in polymer design for scaffolds, mechanical properties are more widely characterized and are relevant for this application. Thus, existing datasets which contain a broader scope of polymers could be adapted to ML approaches for polymer scaffold design.

Biologic sensing. Polymers have been used in biomedical sensing applications as both polymer electronics and encapsulations of traditional electronic sensors[106–108]. As the former, polymeric systems have been developed for strain and pressure sensing in applications such as cardiac monitoring and intracranial pressure. Although deep learning algorithms (e.g., CNN, HMM) have been applied to sense outputs such as electrocardiogram data, there have not been systematic studies relating material composition and device shape to targets, such as sensitivity to external stimuli and conductivity[109]. As encapsulations, properties such as inertness, water absorption, and water barrier performance will greatly benefit from ML prediction and, ultimately, the inverse design of novel polymer systems with these properties. More widely characterized properties like conductivity and mechanical properties are also relevant for sensing applications. Thus, as with biomedical scaffolds, existing datasets may be adaptable for this application.

Challenges and future directions. Current datasets for medically relevant parameters, such as degradation time and water uptake, are small, often exhibit non-standardized methods of characterization and have not been publicly assembled. Thus, the extension of ML methods to designing medically relevant polymers will require a paradigm shift in data curation. More intentional data curation by journals or data sharing by researchers can expand the possibilities of ML with medically relevant polymer synthesis without having to extract data by web scraping, which remains an ethically ambiguous task[110–112]. CRIPT is one promising data sharing platform to this end pending researchers' engagement. The development of molecular dynamic simulation methods and high-throughput synthesis and characterization would make applying ML to biomedical polymers possible within individual groups. Standardization with respect to characterization, data analysis and data presentation of biomedically important properties, such as biocompatibility and degradation time, would also make accumulating larger datasets more feasible.

In addition to data availability concerns, more experimental validation of ML tasks would help establish the utility of these approaches in a wet-lab environment as well as provide insight into the performance of the model. Current work in polymer chemistry more broadly has focused on widely available properties (e.g., thermomechanical properties) and properties which are easily simulated (e.g., bandgap)—some of which are relevant to biomedical applications.

Extending this existing work to new fields is an achievable way to make progress toward expediting polymeric biomaterial design. Ultimately, applied ML represents an extremely promising tool toward accessing state-of-the-art of medically relevant polymers. This reality can be accelerated through:

I. Widespread contribution by experimentalists to data sharing platforms like CRIPT.
II. Standardization of the characterization of biomedically relevant properties (e.g., degradation, water uptake).
III. Development of affordable, high-throughput methods for polymerization via step-growth mechanisms (i.e., the majority of degradable polymers and minority of high-throughput approaches).
IV. Incorporation of coding proficiency and an introduction to ML methods as a part of chemistry curriculum.
V. Collaboration with ML experts and integration of ML methods into current research efforts.

References

1. Gomez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
2. Polykovskiy, D. et al. Entangled conditional adversarial autoencoder for de novo drug discovery. *Mol. Pharm.* **15**, 4398–4405 (2018).
3. Nguyen, D. H. & Tsuda, K. Generating reaction trees with cascaded variational autoencoders. *J. Chem. Phys.* **156**, 044117 (2022).
4. Cai, C. et al. Transfer learning for drug discovery. *J. Med Chem.* **63**, 8683–8694 (2020).
5. Panteleev, J., Gao, H. & Jia, L. Recent applications of machine learning in medicinal chemistry. *Bioorg. Med. Chem. Lett.* **28**, 2807–2815 (2018).
6. Bostrom, J., Brown, D. G., Young, R. J. & Keseru, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discov.* **17**, 709–727 (2018).
7. Schneider, G. & Clark, D. E. Automated de novo drug design: are we nearly there yet? *Angew. Chem. Int Ed. Engl.* **58**, 10792–10803 (2019).
8. Research, G. V. Medical Polymer Market Size, Share & Trends Analysis Report By Product, By Application (Medical Device Packaging, Tooth Implants, Wound Care, Mobility Aids, Denture-based Materials), By Region, And Segment Forecasts, 2022–2030. (Grand View Research).
9. Kerner, J., Dogan, A. & von Recum, H. Machine learning and big data provide crucial insight for future biomaterials discovery and research. *Acta Biomater.* **130**, 54–65 (2021).
10. Turek, P., Budzik, G., Oleksy, M. & Bulanda, K. Polymer materials used in medicine processed by additive techniques. *Polimery* **65**, 510–515 (2020).
11. Roy, N. K., Potter, W. D. & Landau, D. P. Polymer property prediction and optimization using neural networks. *IEEE Trans. Neural Netw.* **17**, 1001–1014 (2006).
12. Wnek, G. E. Structure–Property relationships of small organic molecules as a prelude to the teaching of polymer science. *J. Chem. Educ.* **94**, 1647–1654 (2017).
13. Cencer, M. M., Moore, J. S. & Assary, R. S. Machine learning for polymeric materials: an introduction. *Polymer International*, https://doi.org/10.1002/pi.6345 (2021).
14. Wood, C. in *CNBC (© 2022 CNBC LLC., TECHNOLOGY EXECUTIVE COUNCIL*, 2020).
15. Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
16. Yang, Z. et al. Deep learning approaches for mining structure-property linkages in high contrast composites from simulation datasets. *Comput. Mater. Sci.* **151**, 278–287 (2018).
17. Jung, J., Yoon, J. I., Park, H. K., Kim, J. Y. & Kim, H. S. An efficient machine learning approach to establish structure-property linkages. *Comput. Mater. Sci.* **156**, 17–25 (2019).
18. Chen, C. H., Tanaka, K. & Funatsu, K. Random forest model with combined features: a practical approach to predict liquid-crystalline property. *Mol. Inf.* **38**, e1800095 (2019).
19. Martienssen, W. *Handbook of Materials Data* 2nd edn (Springer, 2018).
20. Mark, J. E. *Polymer Data Handbook* (Oxford University Press, 1998).
21. Holding, S. Polymers: a property database. *Chromatographia* **72**, 587–587 (2010).
22. Huan, T. D. et al. A polymer dataset for accelerated property prediction and design. *Sci. Data* **3**, 160012 (2016).
23. Group, R. R. *Khazana: A Computational Materials Knowledgebase*, https://khazana.gatech.edu/module_search/search.php?m=2 (2022).
24. Jorgensen, P. B. et al. Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, 241735 (2018).
25. Batra, R. et al. Polymers for extreme conditions designed using syntax-directed variational autoencoders. *Chem. Mater.* **32**, 10489–10500 (2020).
26. St John, P. C. et al. Message-passing neural networks for high-throughput polymer screening. *J. Chem. Phys.* **150**, 234111 (2019).
27. Zhan, X. et al. Synthesis, characterization and molecular dynamics simulation of the polyacrylates membranes. *e-Polym.* **16**, 83–89 (2016).
28. Roy, J. K., Pinto, H. P. & Leszczynski, J. Interaction of epoxy-based hydrogels and water: a molecular dynamics simulation study. *J. Mol. Graph Model* **106**, 107915 (2021).
29. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2021).
30. Moore, G. J., Bardagot, O. & Banerji, N. Deep transfer learning: a fast and accurate tool to predict the energy levels of donor molecules for organic photovoltaics. *Adv. Theor. Simul.* **5**, https://doi.org/10.1002/adts.202100511 (2022).
31. Zhang, Y. et al. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Org. Chem. Front.* **8**, 1415–1423 (2021).
32. Yamada, H. et al. Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
33. Wu, S. et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput. Mater.* **5**, https://doi.org/10.1038/s41524-019-0203-2 (2019).
34. Kumar, J. N., Li, Q. & Jun, Y. Challenges and opportunities of polymer design with machine learning and high throughput experimentation. *MRS Commun.* **9**, 537–544 (2019).
35. Gormley, A. J. & Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nat. Rev. Mater.* **6**, 642–644 (2021).
36. Reis, M. et al. Machine-learning-guided discovery of (19)F MRI agents enabled by automated copolymer synthesis. *J. Am. Chem. Soc.* **143**, 17677–17689 (2021).
37. Judzewitsch, P. R., Zhao, L., Wong, E. H. H. & Boyer, C. High-throughput synthesis of antimicrobial copolymers and rapid evaluation of their bioactivity. *Macromolecules* **52**, 3975–3986 (2019).
38. Lin, B., Hedrick, J. L., Park, N. H. & Waymouth, R. M. Programmable high-throughput platform for the rapid and scalable synthesis of polyester and polycarbonate libraries. *J. Am. Chem. Soc.* **141**, 8921–8927 (2019).
39. Upadhya, R. et al. PET-RAFT and SAXS: high throughput tools to study compactness and flexibility of single-chain polymer nanoparticles. *Macromolecules* **52**, 8295–8304 (2019).
40. Zheng, Y., Luo, Y., Feng, K., Zhang, W. & Chen, G. High throughput screening of glycopolymers: balance between cytotoxicity and antibacterial property. *ACS Macro Lett.* **8**, 326–330 (2019).
41. Judzewitsch, P. R. et al. High-throughput process for the discovery of antimicrobial polymers and their upscaled production via flow polymerization. *Macromolecules* **53**, 631–639 (2020).
42. Mann, J. L. et al. An ultrafast insulin formulation enabled by high-throughput screening of engineered polymeric excipients. *Sci. Transl. Med.* **12**, eaba6676 (2020).
43. Tamasi, M., Kosuri, S., DiStefano, J., Chapman, R. & Gormley, A. J. Automation of controlled/living radical polymerization. *Adv. Intell. Syst.* **2**, 1900126 (2020).
44. Baudis, S. & Behl, M. High-throughput and combinatorial approaches for the development of multifunctional polymers. *Macromol. Rapid Commun.* **43**, e2100400 (2022).
45. Rizkin, B. A., Shkolnik, A. S., Ferraro, N. J. & Hartman, R. L. Combining automated microfluidic experimentation with machine

learning for efficient polymerization design. *Nat. Mach. Intell.* **2**, 200–209 (2020).
46. Gurnani, P. et al. PCR-RAFT: rapid high throughput oxygen tolerant RAFT polymer synthesis in a biology laboratory. *Polym. Chem.* **11**, 1230–1236 (2020).
47. Behl, M., Balk, M., Lützow, K. & Lendlein, A. Impact of block sequence on the phase morphology of multiblock copolymers obtained by high-throughput robotic synthesis. *Eur. Polymer J.* **143**, https://doi.org/10.1016/j.eurpolymj.2020.110207 (2021).
48. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).
49. Jablonka, K. M., Jothiappan, G. M., Wang, S., Smit, B. & Yoo, B. Bias free multiobjective active learning for materials design and discovery. *Nat. Commun.* **12**, 2312 (2021).
50. Storti, G. & Lattuada, M. In *Bioresorbable Polymers for Biomedical Applications*: From Fundamentals to Translational Medicine (eds. Perale, G. & Hilboren, J.) 153–179 (2017).
51. Singhvi, M. S., Zinjarde, S. S. & Gokhale, D. V. Polylactic acid: synthesis and biomedical applications. *J. Appl. Microbiol* **127**, 1612–1626 (2019).
52. Das, A. & Mahanwar, P. A brief discussion on advances in polyurethane applications. *Adv. Ind. Eng. Polym. Res.* **3**, 93–101 (2020).
53. Shakiba, M. et al. Nylon—A material introduction and overview for biomedical applications. *Polym. Adv. Technol.* **32**, 3368–3383 (2021).
54. Akinc, A., Lynn, D. M., Anderson, D. G. & Langer, R. Parallel synthesis and biophysical characterization of a degradable polymer library for gene delivery. *J. Am. Chem. Soc.* **125**, 5316–5323 (2003).
55. Walsha, D. J. et al. CRIPT: a scalable polymer material data structure. *ChemRxiv*, https://doi.org/10.26434/chemrxiv-2022-xpz37 (2022).
56. Deagen, M. E. et al. FAIR and interactive data graphics from a scientific knowledge graph. *Sci. Data* **9**, 239 (2022).
57. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
58. Drefahl, A. CurlySMILES: a chemical language to customize and annotate encodings of molecular and nanodevice structures. *J. Cheminform.* **3**, 1 (2011).
59. Lin, T. S. et al. BigSMILES: a structurally-based line notation for describing macromolecules. *ACS Cent. Sci.* **5**, 1523–1531 (2019).
60. Chen, G., Tao, L. & Li, Y. Predicting polymers' glass transition temperature by a chemical language processing model. *Polymers (Basel)* **13**, 1898 (2021).
61. Patel, R. A., Borca, C. H. & Webb, M. A. Featurization strategies for polymer sequence or composition design by machine learning. *Mol. Syst. Des. Eng.* **7**, 661–676 (2022).
62. Aldeghi, M. & Coley, C. W. A graph representation of molecular ensembles for polymer property prediction. *Chem. Sci.* **13**, 10486–10498 (2022).
63. Mentel, L. et al. *Awesome Python Chemistry*, https://github.com/lmmentel/awesome-python-chemistry#machine-learning (2022).
64. Huang, Y. et al. Structure–Property correlation study for organic photovoltaic polymer materials using data science approach. *J. Phys. Chem. C* **124**, 12871–12882 (2020).
65. Meyer, T. A., Ramirez, C., Tamasi, M. J. & Gormley, A. J. A user's guide to machine learning for polymeric biomaterials. *ACS Polym. Au* **17**, 141–157 (2022).
66. Ghosh, S. & Dasgupta, R. in *Machine Learning in Biological Sciences: Updates and Future Prospects* (eds Shyamasree Ghosh & Rathi Dasgupta) 51–57 (Springer Nature Singapore, 2022).
67. Suwardi, A. et al. Machine learning-driven biomaterials evolution. *Adv. Mater.* **34**, e2102703 (2022).
68. Bradford, G. et al. Chemistry-informed machine learning for polymer electrolyte discovery. *ACS Cent. Sci.* **9**, 206–216 (2023).
69. Karniadakis, G. E. et al. Physics-informed machine learning. *Nat. Rev. Phys.* **3**, 422–440 (2021).
70. Oviedo, F., Ferres, J. L., Buonassisi, T. & Butler, K. T. Interpretable and explainable machine learning for materials science and chemistry. *Acc. Mater. Res.* **3**, 597–607 (2022).
71. Rudin, C. et al. Interpretable machine learning: fundamental principles and 10 grand challenges. *Stat. Surv.* **16**, 1–85 (2022).
72. Gramegna, A. & Giudici, P. SHAP and LIME: an evaluation of discriminative power in credit risk. *Front Artif. Intell.* **4**, 752558 (2021).
73. Kumar, R., Le, N., Oviedo, F., Brown, M. E. & Reineke, T. M. Combinatorial polycation synthesis and causal machine learning reveal divergent polymer design rules for effective pDNA and ribonucleoprotein delivery. *JACS Au* **2**, 428–442 (2022).
74. Kaddour, J., Lynch, A., Liu, Q., Kusner, M. J. & Silva, R. Causal machine learning: a survey and open problems. *arXiv*, arXiv:2206.15475v2 (2022).
75. Sherman, Z. M., Howard, M. P., Lindquist, B. A., Jadrich, R. B. & Truskett, T. M. Inverse methods for design of soft materials. *J. Chem. Phys.* **152**, 140902 (2020).
76. Sattari, K., Xie, Y. & Lin, J. Data-driven algorithms for inverse design of polymers. *Soft Matter* **17**, 7607–7622 (2021).
77. Patra, T. K., Loeffler, T. D. & Sankaranarayanan, S. Accelerating copolymer inverse design using monte carlo tree search. *Nanoscale* **12**, 23653–23662 (2020).
78. Kim, C., Batra, R., Chen, L., Tran, H. & Ramprasad, R. Polymer design using genetic algorithm and machine learning. *Comput. Mater. Sci.* **186**, https://doi.org/10.1016/j.commatsci.2020.110067 (2021).
79. Kumar, J. N. et al. Machine learning enables polymer cloud-point engineering via inverse design. *npj Comput. Mater.* **5**, https://doi.org/10.1038/s41524-019-0209-9 (2019).
80. Sanchez-Lengeling, B. & Aspuru-Guzik, A. N. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
81. Santana, R. et al. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. *Nanoscale* **12**, 13471–13483 (2020).
82. Bannigan, P. et al. Machine learning models to accelerate the design of polymeric long-acting injectables. *ChemRxiv*, https://doi.org/10.26434/chemrxiv-2021-mxrxw-v3 (2022).
83. Caccavo, D. An overview on the mathematical modeling of hydrogels' behavior for drug delivery systems. *Int. J. Pharm.* **560**, 175–190 (2019).
84. Rothstein, S. N. & Little, S. R. A "tool box" for rational design of degradable controlled release formulations. *J. Mater. Chem.* **21**, 29–39 (2011).
85. Perni, S. & Prokopovich, P. Feasibility and application of machine learning enabled fast screening of poly-beta-amino-esters for cartilage therapies. *Sci. Rep.* **12**, 14215 (2022).
86. Mikulskis, P. et al. Prediction of broad-spectrum pathogen attachment to coating materials for biomedical devices. *ACS Appl. Mater. Interfaces* **10**, 139–149 (2018).
87. Le, T. C., Penna, M., Winkler, D. A. & Yarovsky, I. Quantitative design rules for protein-resistant surface coatings using machine learning. *Sci. Rep.* **9**, 265 (2019).
88. Epa, V. C. et al. Modelling and prediction of bacterial attachment to polymers. *Adv. Funct. Mater.* **24**, 2085–2093 (2014).
89. Conev, A. et al. Machine learning-guided three-dimensional printing of tissue engineering scaffolds. *Tissue Eng. Part A* **26**, 1359–1368 (2020).
90. Damiati, S. A. & Damiati, S. Microfluidic synthesis of indomethacin-loaded PLGA microparticles optimized by machine learning. *Front. Mol. Biosci.* **8**, 677547 (2021).

91. Kumar, R. et al. Efficient polymer-mediated delivery of gene-editing ribonucleoprotein payloads through combinatorial design, parallelized experimentation, and machine learning. *ACS Nano* **14**, 17626–17639 (2020).
92. Chan, D. et al. Combinatorial polyacrylamide hydrogels for preventing biofouling on implantable biosensors. *Adv. Mater.* **34**, 2109764 (2022).
93. Tamasi, M. J. et al. Machine learning on a robotic platform for the design of polymer-protein hybrids. *Adv. Mater.* **34**, e2201809 (2022).
94. Ji, Z. et al. Machine learning models for predicting cytotoxicity of nanomaterials. *Chem. Res. Toxicol.* **35**, 125–139 (2022).
95. Xu, J., Lin, X. & Gowen, A. A. Combining machine learning with meta-analysis for predicting cytotoxicity of micro- and nanoplastics. *Journal of Hazardous Materials Advances* **8**, (2022).
96. Ma, Z., Wu, Y., Wang, J. & Liu, C. In vitro and in vivo degradation behavior of poly(trimethylene carbonate-co-d,l-lactic acid) copolymer. *Regen. Biomater.* **4**, 207–213 (2017).
97. Pappalardo, D., Mathisen, T. & Finne-Wistrand, A. Biocompatibility of resorbable polymers: a historical perspective and framework for the future. *Biomacromolecules* **20**, 1465–1477 (2019).
98. Arun, Y., Ghosh, R. & Domb, A. J. Biodegradable hydrophobic injectable polymers for drug delivery and regenerative medicine. *Adv. Funct. Mater.* **31**, https://doi.org/10.1002/adfm.202010284 (2021).
99. Kavand, A., Anton, N., Vandamme, T., Serra, C. A. & Chan-Seng, D. Synthesis and functionalization of hyperbranched polymers for targeted drug delivery. *J. Control Release* **321**, 285–311 (2020).
100. Braatz, D. et al. Chemical approaches to synthetic drug delivery systems for systemic applications. *Angew. Chem. Int. Ed. Engl.* **61**, e202203942 (2022).
101. Owh, C., Ho, D., Loh, X. J. & Xue, K. Towards machine learning for hydrogel drug delivery systems. *Trends Biotechnol.* https://doi.org/10.1016/j.tibtech.2022.09.019 (2022).
102. Ohnsorg, M. L. et al. Bottlebrush polymer excipients enhance drug solubility: influence of end-group hydrophilicity and thermoresponsiveness. *ACS Macro Lett.* **10**, 375–381 (2021).
103. Freeman, S., Calabro, S., Williams, R., Jin, S. & Ye, K. Bioink formulation and machine learning-empowered bioprinting optimization. *Front. Bioeng. Biotechnol.* **10**, 913579 (2022).
104. Motta, C. M. M., Endres, K. J., Wesdemiotis, C., Willits, R. K. & Becker, M. L. Enhancing Schwann cell migration using concentration gradients of laminin-derived peptides. *Biomaterials* **218**, 119335 (2019).
105. Kirillova, A. et al. Fabrication of biomedical scaffolds using biodegradable polymers. *Chem. Rev.* **121**, 11238–11304, https://doi.org/10.1021/acs.chemrev.0c01200 (2021).
106. Liu, Y., Feig, V. R. & Bao, Z. Conjugated polymer for implantable electronics toward clinical application. *Adv. Health. Mater.* **10**, e2001916 (2021).
107. Choi, Y. S. et al. Biodegradable polyanhydrides as encapsulation layers for transient electronics. *Adv. Funct. Mater.* **30**, https://doi.org/10.1002/adfm.202000941 (2020).
108. Zeglio, E., Rutz, A. L., Winkler, T. E., Malliaras, G. G. & Herland, A. Conjugated polymers for assessing and controlling biological functions. *Adv. Mater.* **31**, e1806712 (2019).
109. Kwon, S. H. & Dong, L. Flexible sensors and machine learning for heart monitoring. *Nano Energy* **102**, https://doi.org/10.1016/j.nanoen.2022.107632 (2022).
110. Thomas, D. M. & Mathur, S. in *3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)* 450–454 (Coimbatore, India, 2019).
111. Breuer, J., Bishop, L. & Kinder-Kurlanda, K. The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships. *N. Media Soc.* **22**, 2058–2080 (2020).
112. Riley, K. C. Data scraping as a cause of action: limiting use of the CFAA and trespass in online copying cases. *Fordham Intellect. Prop. Media Entertain. Law J.* **29** (2018).
113. Otsuka, S., Kuwajima, I., Hosoya, J., Xu, Y. & Yamazaki, M. in *2011 International Conference on Emerging Intelligent Data and Web Technologies*. 22–29.
114. *Polymers: A Property Database 2021*, https://poly.chemnetbase.com/faces/polymers/PolymerSearch.xhtml (2021).
115. *Polymer Property Predictor and Database*, https://pppdb.uchicago.edu (Center for Hierarchical Materials Design (CHiMaD), 2022).
116. *MatWeb: Material Property Data*, https://www.matweb.com/index.aspx (2022).
117. Rebello, N. J. et al. *Block Copolymer Phase Behavior Database (BCDB)*, https://github.com/olsenlabmit/BCDB (2021).
118. Dabrowski-Tumanski, P., Rubach, P., Niemyska, W., Gren, B. A. & Sulkowska, J. I. Topoly: python package to analyze topology of polymers. *Brief. Bioinform.* **22**, bbaa196 (2021).
119. Sahu, H., Shen, K. H., Montoya, J. H., Tran, H. & Ramprasad, R. Polymer structure predictor (PSP): a python toolkit for predicting atomic-level structural models for a range of polymer geometries. *J. Chem. Theory Comput* **18**, 2737–2748 (2022).
120. Santana-Bonilla, A. & Lorenz, C. *PySoftK*, https://github.com/alejandrosantanabonilla/pysoftk (2022).
121. Fortunato, M. E. & Colina, C. M. pysimm: a python package for simulation of molecular systems. *SoftwareX* **6**, 7–12 (2017).
122. Grunewald, F. et al. Polyply; a python suite for facilitating simulations of macromolecules and nanomaterials. *Nat. Commun.* **13**, 68 (2022).
123. Wilson, N. St., John, P. & Crowley, M. m2p (Monomers to Polymers). *USDOE Office of Energy Efficiency and Renewable Energy (EERE), Transportation Office*, https://doi.org/10.11578/dc.20200922.9 (2020).
124. Shen, K.-H. K. & Group, R. *Polymer Molecular Dynamics toolkit*, https://github.com/Ramprasad-Group/Polymer-Molecular-Dynamics (2022).
125. *BigSMILES_parser*, https://github.com/tzyyshyanglin/BigSMILES_parser (2019).
126. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
127. Yamano, H. et al. Predicting and considering properties of general polymers using incomplete dataset in *2020 International Symposium on Semiconductor Manufacturing (ISSM)* 1–3 (Institute of Electrical and Electronics Engineers (IEEE), Tokyo, Japan, 2020).
128. Lee, F. L. et al. Comparison of machine learning methods towards developing interpretable polyamide property prediction. *Polymers (Basel)* **13**, 3653 (2021).
129. Duce, C., Micheli, A., Starita, A., Tiné, M. R. & Solaro, R. Prediction of polymer properties from their structure by recursive neural networks. *Macromol. Rapid Commun.* **27**, 711–715 (2006).
130. Dennis, J. M. & Zubarev, D. Y. Hebbian learning on small data enables experimental discovery of high Tg polyimides. *J. Phys. Chem. A* **125**, 6829–6835 (2021).
131. Jha, A., Chandrasekaran, A., Kim, C. & Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Model. Simul. Mater. Sci. Eng.* **27**, https://doi.org/10.1088/1361-651X/aaf8ca (2019).
132. Pilania, G., Iverson, C. N., Lookman, T. & Marrone, B. L. Machine-learning-based predictive modeling of glass transition temperatures: a case of polyhydroxyalkanoate homopolymers and copolymers. *J. Chem. Inf. Model* **59**, 5013–5025 (2019).

133. Pugar, J. A., Childs, C. M., Huang, C., Haider, K. W. & Washburn, N. R. Elucidating the physicochemical basis of the glass transition temperature in linear polyurethane elastomers with machine learning. *J. Phys. Chem. B* **124**, 9722–9733 (2020).
134. Tao, L., Varshney, V. & Li, Y. Benchmarking machine learning models for polymer informatics: an example of glass transition temperature. *J. Chem. Inf. Model* **61**, 5395–5413 (2021).
135. Nazarova, A. L. et al. Dielectric polymer property prediction using recurrent neural networks with optimizations. *J. Chem. Inf. Model* **61**, 2175–2186 (2021).
136. Maouz, H. et al. QSPR studije karbonilnih, hidroksilnih, polienskih indeksa i prosječne molekulske težine polimera pod fotostabilizacijom pristupom ANN i MLR. *Kem. u. Ind.* **69**, 1–16 (2020).
137. Afzal, M. A. F., Cheng, C. & Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *J. Chem. Phys.* **148**, 241712 (2018).
138. Barnett, J. W. et al. Designing exceptional gas-separation polymer membranes using machine learning. *Sci. Adv.* **6**, eaaz4301 (2020).

Acknowledgements

S.M. and Q.L. thank the National Science Foundation Artificial Intelligence for Designing and Understanding Materials—National Research Traineeship (aiM-NRT) at Duke University for funding this effort under grant DGE-2022040 as well as domain knowledge in machine learning and material science. M.L.B. is grateful for financial support from a National Science Foundation - FDA Scholar in Residence Award (CBET 2129615).

Author contributions

All authors made critical changes to the manuscript and contributed to its actualization.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Matthew L. Becker.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work.