Simulation and Machine Learning Prediction of Clogging Processes in Porous Media

Francisco J. Carrillo*
Department of Chemical and Biological Engineering, Princeton University, Princeton, NJ, USA

Ian C. Bourg†
Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA
and
High Meadows Environmental Institute, Princeton University, Princeton, NJ, USA

(Dated: September 6, 2021)

The accumulation of discrete masses within constrained flow conduits is an common phenomenon within both natural and industrial settings: it describes the clogging of pipes, roads, oil reservoirs, rivers, and arteries. In this study we use Computational Fluid Mechanics and Discrete Element Models to run over 2,000 different clogging simulations in randomly-generated porous media in order to train and evaluate the clogging prediction performance of several Machine Learning algorithms. The resulting best-performing classifier (an Extremely Randomized Trees algorithm) is able to predict clogging *a-priori* with an accuracy of 0.96 and 0.91 in numerical and experimental systems, respectively. Similarly, the best performing regressor (also a decision tree-based algorithm) is able to achieve an R^2 value of 0.93 when predicting the degree of clogging in said systems. We believe this standardized computational tool has the potential to help evaluate the design process of engineered and natural porous media.

I. INTRODUCTION

The erosion, transport, and deposition of suspended particles is ubiquitous within both natural and engineered porous systems. These coupled processes control the evolution of sedimentary formations, the distribution of contaminants in the environment, and the clogging of porous media, pipes, and arteries [1–4]. Although material transport in porous media is relatively well understood, the study of particle deposition and accumulation is still in its infancy.

The difficulty of predicting clogging in porous media arises from the inherently complex nature of the relevant systems. Clogging in porous media is notoriously hard to probe and characterize, as it is necessary to account for each system's 3D geometry, grain size distribution, pore size distribution, rock/fluid chemistry, particle size, and even surface charge [5–10]. Added to these spatial challenges is the fact that clogging is also a temporal process. Clogging mechanisms often don't reach a steady state: Clogs can form, redirect fluid fields, change pressure gradients, break, and form again downstream [11].

Experimentalists have taken two separate routes in addressing these challenges. The first approach attempts to characterize particle deposition and clogging by reducing the complexity of the studied systems through the creation of simplified micro-models [12–15]. This allows a measure of control of a handful of relevant variables (fluid flow rate, particle size, pore size, flow time) while significantly constraining others (flow path complexity and material heterogeneity). These studies have yielded noteworthy results. In particular, they have shown that clogging scales as a function of the pore to particle size ratio [8] and that, under certain conditions, it is independent of particle injection rate and system porosity [12]. However, it is not clear if these conclusions still hold for more complex natural systems.

The second approach relies on using advanced imaging techniques to probe real porous systems (or close approximations to them). These often rely on computed X-ray micro-tomography (XCT) and confocal microscopy. The former has taken a key role in characterizing particle aggregate formation in natural rock samples and identifying their dependence on fluid chemistry and rock geometry [7, 16–18]. However, although highly informative, this technique suffers from not having large enough fields-of-view or high enough temporal resolutions. Confocal microscopy takes advantage of the fact that it is possible to match the refractive index of an artificial porous medium with the index of its permeating fluid phase in order to "see" through the whole solid-fluid mixture. This technique has allowed scientists to characterize particle deposition and clogging as a function of time while having a high level of control of the relevant flow variables and material characteristics [11, 19, 20].

However, all experimental studies suffer from the same practical problem: they cannot probe a large enough parametric phase-space as required to properly generalize their results to naturally-occurring porous media due to the difficulty of creating and running an experiment with more than 2-3 independent variables. Historically, experimentalists have studied clogging by varying the amount of particles flowing through the porous medium, the fluid flow velocity[12, 13], inter-particle interactions [3, 7, 21], particle-wall interactions [10, 16], particle size distribution [2], and/or pore size distribution [5, 8].

* https://github.com/Franjcf
† http://bourg.princeton.edu

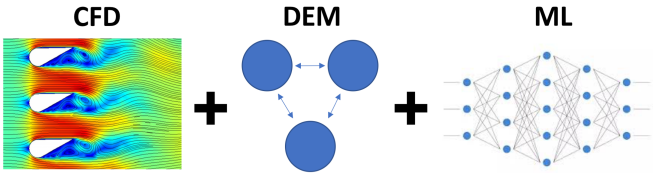

FIG. 1. Conceptual representation of the models we used for clogging simulation and prediction.

The necessity to investigate a large range of systems is evidenced by the fact that experimental studies often reach conflicting conclusions. For example, some studies maintain that particle transport suppresses fluid flow by reducing the permeability of porous media [7, 22, 23], while others state that particles can actually help enhance flow [20, 24–26]. Even so, assuming that we could obtain a large-enough data set, there is no guarantee that we could derive general relationships between variables: some conclusions may only hold at low flow rates and/or low grain heterogeneity, while others may only hold at low fluid viscosities and/or high particle concentrations. One way to address this problem is through the application of Machine Learning (ML) algorithms to analyze such large data sets. These approaches have been shown to learn and predict complex relations in a wide variety of fields: from predicting fluid turbulence in jet engines [27] to enabling image recognition in autonomous vehicles [28]. These models, however, require highly extensive data sets to become effective predictors.

Fortunately, due to the rise of large-scale parallelized computing, it is now possible to run thousands of particle transport simulations in order to systematically study 10-15 dimensional parametric phase spaces. However, to the best of our knowledge, there have not been any studies that have used numerical models for this purpose in the context of particle flow through porous media. Current popular approaches rely on continuum-level (i.e. volume averaging) approximations such as filtration theory to study particle deposition [1, 29, 30] and erosion [31, 32]. However, these relatively fast models require phenomenological kinetic parameters for proper parameterization, meaning they are often more descriptive than predictive. Coupled CFD-DEM approaches constitute a popular alternative [4, 9, 33–36]. In these models particle deposition is simulated through careful consideration of coupled particle-fluid mechanics and direct numerical simulations. However, due to their complexity and relatively-high computational cost, very few studies have attempted to model particle mechanics through these means.

In this paper, we leverage Princeton's University's large computational resources in order to explore the underlying principles that control clogging mechanics. In particular, we explore the feasibility of using ML approaches to create a computational tool that can predict clogging *a-priori* and that can be used to improve and streamline the design process of engineered porous systems. To do so, we ran over 2000 CFD-DEM simulations in randomly-generated porous geometries while systematically varying 12 of the system's design parameters (pore size, pore size heterogeneity, grain size heterogeneity, particle diameter, particle-particle attraction, particle-wall attraction, particle flux, porosity, system size, fluid velocity, and fluid viscosity). The resulting cases where then labeled and analyzed in order to train and evaluate several different ML classifiers and regressors, which we then used to predict clogging in equivalent numerical systems. Then, through model optimization, we identified which training features were most indicative of clogging in heterogeneous porous media. Lastly, we validated our methodology by predicting clogging in real experimental systems obtained from three separate studies. To the best of our knowledge, this is the first time ML approaches have been successfully used to reliably predict and control clogging processes.

II. METHODS

A. Coupled Fluid and Particle Flow Mechanics

Our numerical simulations where performed on the CFDEM® computational framework, which couples Computational Fluid Dynamics (CFD) simulations performed in OpenFOAM® with granular particle flow simulations performed in LIGGGHTS®. Both C++ libraries are free-to-use, open-source, parallelizable, and easily customizable platforms.

In our case, fluid flow was simulated by applying the Finite Volume Method to discretize and solve the Navier-Stokes mass and momentum conservation equations. In turn, we used LIGGGHTS® to describe the motion and interaction of a large number of individual particles within a grid-free environment through Lagragian Discrete Element Models (DEM). Here, particle behaviour is dictated by the addition of all the hydrodynamic forces, contact particle-particle/wall forces, and attraction/repulsion forces acting on each individual element.

CFDEM® couples the fluid and particle models through an iterative algorithm that super-imposes both models' physical domains. To do so, an additional particle-drag term is added to the fluid's momentum conservation equation to account for particle movement, and an additional fluid-drag term is added to each particle's force balance to account for fluid-induced motion. Each model can probe its counterpart at any given moment in time, at which point the CFDEM® algorithm iterates over both solver's solutions until convergence is reached. Please refer to [33, 34] and [35] for an in depth explanation and verification of the underlying models. Our specific numerical setup will be further discussed in Section II C.

B. Creating Randomized Porous Systems

The coupled fluid and particle mechanics simulations where performed on a set of randomly-generated porous geometries. These porous systems were created through a custom Matlab® script which allowed us to specify the following criteria: average grain size, grain size standard deviation, average pore size (i.e. shortest distance between two given grains), pore size standard deviation, and porosity (see Figure 2). Note that, for simplicity, all grains where assumed to be cylindrical in shape and the domain's length and width where kept constant at 50 cm. Given a fixed domain space, the porosity and pore size distribution become coupled variables, which is why only one of these could be specified for a given geometry (the other one was measured at the end of geometry generation). The result is an algorithm that can generate well-characterized, yet randomized porous systems on which we can perform our numerical simulations. Finally, we note that special care was taken during the mesh generation process in order to make sure each pore was resolved by a minimum of 10 grid cells and where the minimum cross-sectional grid resolution was 500 by 500 cells.

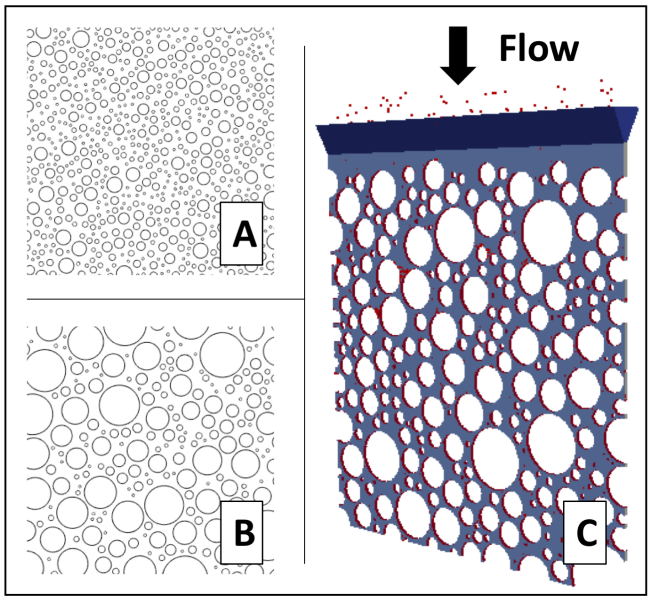

FIG. 2. A,B) Examples of randomly-generated porous media configurations with different grain sizes and porosities. C) Example of the numerical mesh and simulation setup. Here, the areas within the cylinders (i.e. grains) are removed during meshing to making them inaccessible to fluids and particles. Also note the direction of flow and the fact that the only inlets and outlets are at the upper and lower boundaries of the geometry, respectively.

C. Simulation Setup

As discussed earlier, the main advantage that numerical simulations have over conventional experiments is their ability to probe large parameter spaces with ease. As such, we ran over 2000 uniquely-parameterized coupled fluid-particle simulations by systematically varying each system's porosity (ϕ), average pore size (P), pore size standard deviation (P_{std}), average grain size (G), grain size standard deviation (G_{std}), geometric thickness of the porous medium (T), particle diameter (D), particle flux (F), particle-particle attraction (PP), particle-wall attraction (PW), fluid velocity (U), and fluid kinematic viscosity (ν). The ranges over which these 12 parameters were varied are shown in Table II.

All other parameters were kept constant. In particular, the fluid (i.e. water) density was 1000 kg/m^3, particle density was 1200 kg/m^3, and gravity ($g = 9.8m/s^2$) was set to be constant in the direction of flow. In order to avoid modeling trivial clogging cases dictated by simple size exclusion, we exclusively considered systems where the particles' diameter was less than the pore size and the geometric thickness (see Figure 3). Said particles were introduced into the system through random non-overlapping seeding at the inlet. Each simulation was set to run for 5 hydraulic residence times (0.5 m / fluid velocity), which took on average about 5.5 hours of computing time on a 28-core Broadwell Xeon node.

Lastly, particle-particle/wall attractions were modeled through the SJKR model, particle-particle/wall collisions where captured through the Granular Hertz model with a Poisson Ratio of 0.22, and the two-way fluid-particle drag coupling was calculated through the DiFelice Drag model every 0.001 seconds [37]. These models imply the following: A) Attraction forces are normal to the inter-

Property	Range
Porosity (ϕ)	0.26-0.98
Average pore size (P)	10^{-4}-10^{-2} m
Pore size std. (P_{std})	5×10^{-4}-5×10^{-3} m
Average grain size (G)	10^{-3}-5×10^{-2} m
Grain size std. (G_{std})	10^{-3}-10^{-2} m
Geometric thickness (T)	10^{-4}-10^{-2} m
Particle diameter (D)	10^{-4}-10^{-2} m
Particle flux (F)	10^{2}-10^{4} particles/s
Particle-particle attraction (PP)	0-5×10^{6} J/m^3
Particle-wall attraction (PW)	0-5×10^{6} J/m^3
Fluid velocity (U)	0-1.5 m/s
Fluid kinematic viscosity (ν)	10^{-2}-10^{-8} m^2/s.

TABLE I. Table of Varied Fluid, Particle, and Structural Properties. Note "std" refers to the standard deviation.

particle contact area and dominate once particles are in direct contact with each other. B) Inter-particle collisions are modeled by defining both a normal force (spring + damping forces) and a tangential force (shear + damping forces). This approach allows particles to both bounce and roll around obstacles and other particles. C) Fluid-particle interactions are governed by a drag force that is proportional to the relative velocity of a given particle and the underlying fluid, the solid volume fraction in the specified control volume (i.e. grid cell), and the square of the Reynolds number [38]. Please refer to [33, 34] and [35] for further discussion of the underlying models.

III. DATA ANALYSIS AND MODEL TRAINING

A. Identifying Clogged Systems

After running all the simulations, it was crucial to accurately and consistently characterize the final configuration of each system in order to properly train and test our ML algorithms. This process was complicated by the fact that clogging in a heterogeneous porous medium is not necessarily a discrete state. As shown Figure 3, clogging might only occur at certain pore throats, might not occur at all, or might occur throughout the porous medium [11]. As such, in order to properly label our clogging geometries for ML classification (or obtain a continuous prediction variable for ML regression) we had to define an objective metric that could characterize the degree of clogging in each system. We called this variable the "Clogging Number" (CN).

The Clogging Number (Eqn. 1) is the product of two independent factors obtained at the end of each simulation : 1) The average relative distance between particles (DP). 2) The symmetry of the particle's final velocity distribution (SYMM). The significance of the first factor is fairly intuitive; systems with particles that end up closer together tend to have a higher degree of clogging. The second factor describes the relative behaviour of all the particles throughout the porous medium, where a high velocity symmetry indicates homogeneous particle behaviour (i.e. the system is either fully clogged or fully unclogged) and a low symmetry indicates heterogeneous behaviour (i.e. some particles are moving and some are stationary). The resulting clogging number can be thought of as an analog the collection efficiency coefficient used in continuum models [2] and is conceptually similar to the clogging density used in Gerber et. al., 2018 [8].

$$\mathrm{CN} = \mathrm{DP} * \mathrm{SYMM} \tag{1}$$

Multiplying both factors together into the CN allows us to characterize each simulation's degree of clogging, where DP tells us if the system tends to be clogged or unclogged, and SYMM tells us the degree to which they are one or the other. After testing several different averaging procedures and symmetry measures, the following two metrics yielded the best descriptive performance. DP was calculated by averaging the average distance between each particle and its closest 50 neighbors, and then non-dimensionalizing against the particle radius. In turn, SYMM was determined by calculating the ratio between the particles' median velocity and their mean velocity. This ratio describes the symmetry of the distribution by using the fact that the median of a sample is independent of changes in its extreme values, while the mean of a sample is. Therefore, whenever a group of particles form a blockage in an otherwise unclogged geometry, the median of the velocity will change at a much slower rate than the mean velocity and their ratio will decrease. Conversely, if we have a fully clogged or fully unclogged system, the velocity distribution will have a higher degree of symmetry, and this ratio will be close to one.

Figure 4 shows that CN provides us with a suitable way of characterizing the final state our clogging simulations. It also gives us a straightforward way to form discrete labels for classification out of this continuous variable, where all samples with CN $\leq$ 1 will be labeled as "Clogged" (1) and all others as "Unclogged" (0). This threshold was chosen by identifying the best linear classifier that could separate a sample of 300 manually-labeled fully clogged and fully unclogged systems (without considering any semi-clogged samples). Applying this threshold to the complete data set yields 1012 samples classified as Clogged and 988 classified as Unclogged.

B. Feature Engineering and Standardization

Having labeled all of our cases, we now proceed to define and scale the training features for our ML algorithms. The objective is to create a standardized set of features that can be readily quantified from any system involving the flow of solid bodies through a set of static obstacles. Scaling the variables also allows us to add information to the model by explicitly dictating important relationships between variables (such as the ratio of particle diameter to pore size). The following training features are based on the varied parameters shown in Table II and can all be obtained *a-priori* at the start of the simulations. As such, a well trained ML model will be able to use the following features to predict clogging without actually having to run any simulations or experiments:
the particle diameter - pore size ratio (D/P)

$$= D/P \tag{2}$$

the particle diameter - geometric thickness ratio (D/T)

$$= D/T \tag{3}$$

the particle diameter - grain size ratio (D/G):

$$= D/G \tag{4}$$

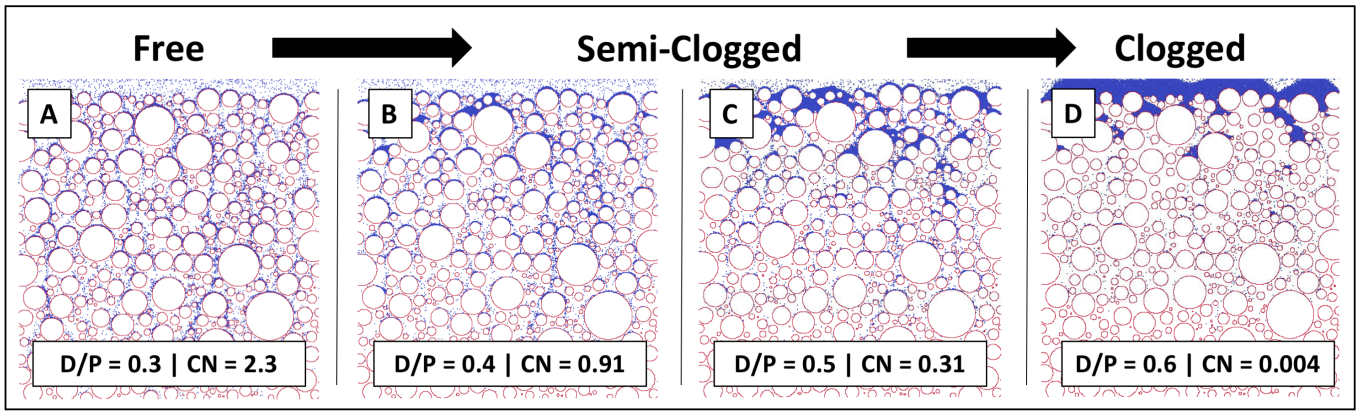

FIG. 3. Examples of different clogging levels. Each panel in this figure shows the final state of four separate clogging simulations with identical geometries and flow conditions. Their only difference is the size of the simulated particles (shown here as the particle diameter to pore size ratio, D/P, which increases from left to right) and their resulting clogging number (CN). A) Free geometry. B, C) Semi-clogged geometries D) Fully-clogged geometry.

the pore size - grain size ratio (P_{nd}):

$$= P/G \tag{5}$$

the geometric thickness - pore size ratio (T/P):

$$= T/P \tag{6}$$

the non-dimensional fluid velocity (U_{nd}):

$$= \frac{U \times \phi \times D}{P \times F} \tag{7}$$

the non-dimensional particle-particle attraction (PP_{nd}):

$$= \frac{PP \times \pi \times (D/2)^2}{(\text{Particle Mass} \times \mathbf{g})} \tag{8}$$

the non-dimensional particle-wall attraction (PW_{nd}):

$$= \frac{PW \times \pi \times (D/2)^2}{(\text{Particle Mass} \times g)} \tag{9}$$

the weighted particle flux (F_w):

$$= F/\phi \tag{10}$$

the logarithm of the kinematic viscosity ($log(\nu)$):

$$= log_{10}(\nu) \tag{11}$$

and the porosity (ϕ). These 11 variables where then used to train the classifiers and the regressors showcased in the following sections.

C. Training and Identifying Best Classifier

The ML classifiers shown in Table II were trained and tested by using k-folds cross-validation (k=5) on all cases, resulting in training sample of $\sim$ 1600 and a testing sample of $\sim$ 400. These classifiers were chosen by virtue of their differing approaches, as we had little-to-no intuition of which one would work best for our purposes (no one has ever used ML to predict clogging mechanisms). Classifier performance was evaluated by calculating their testing accuracy, precision, the area under their Receiving Operating Characteristic (ROC) curve, and quantifying the total number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FP). In this case, a true positive pertains to correctly labeling a case as "Clogged". These metrics where chosen due to their ability to quantify overall classifier performance (accuracy), while also testing for its ability to correctly predict positive clogging labels (precision) and minimize false negatives (ROC). Furthermore, we implemented a grid-search algorithm in order to optimize our choice of model hyperparameters and thus improve the predictive power of each type of classifier. Table II shows the performance metrics of each classifier as a result of average precision grid-search optimization.

The results in Table II are very encouraging, showing that the best classifier, an Extremely Randomized Trees algorithm [39], achieves a classifying accuracy of 0.96, a precision of 0.93, and an ROC of 0.99, meaning that it favors producing false positives over false negatives (a desirable feature in this case). Further analysis into the misclassified samples show that these are partially-clogged samples located close to the classification decision boundary defined by CN = 1 (see Figure 4B). Please refer to the Supplementary Materials for a full description and implementation of the optimized model.

Classifier	Accuracy	Precision	ROC	TP	FP	TN	FN
Bernoulli Naive Bayes	0.53	0.52	0.52	93	82	119	109
Random Forests	0.77	0.71	0.91	147	38	163	55
Logistic Regression	0.78	0.72	0.87	166	51	150	36
Support Vector Machine	0.78	0.72	0.87	165	50	151	37
K-Nearest Neighbors	0.86	0.81	0.94	172	26	175	30
Multilayer Perceptron	0.88	0.83	0.94	180	27	174	22
Extremely Randomized Trees	**0.96**	**0.94**	**0.99**	**198**	**10**	**191**	**4**

TABLE II. Clogging prediction performance of the chosen ML classifiers. The section in bold represent the classifier with the best performance.

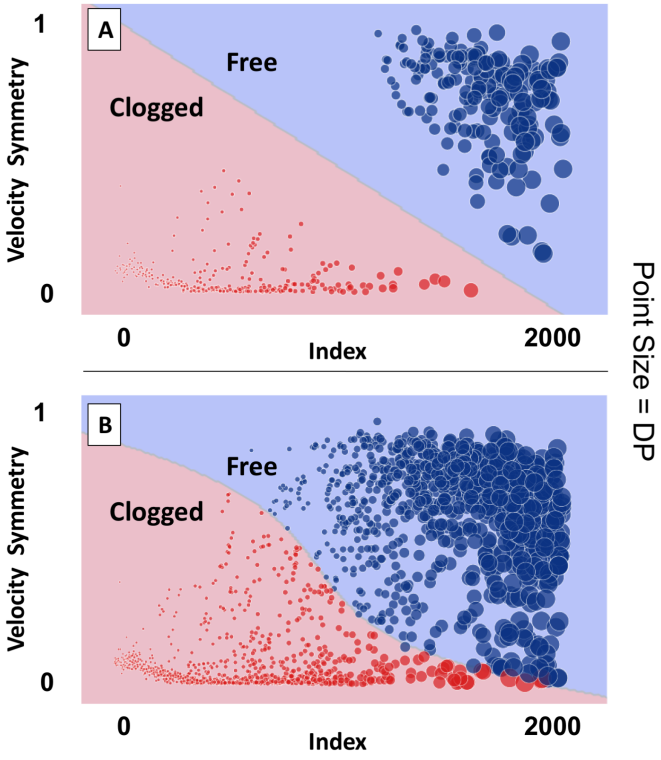

FIG. 4. Clogging sample distribution as a function of the average final distance between particles (DP) and symmetry of the particles' final velocity distribution (SYMM). Here, red and blue colors represent clogged and unclogged samples, respectively. Note that point size is directly proportional to DP and that all cases where plotted from lowest DP to highest DP values for visualization purposes. A) Result of using CN ≤ 1 to separate 300 manually-labelled fully clogged and unclogged samples. B) Result of using CN ≤ 1 to label all data samples. The color contours in both graphs represent the phase-space prediction of the optimized MLP classifier trained with their respective samples and discussed in Section III C.

D. Classification Feature Analysis

Although the testing metrics of our models are relatively high, it might not be practical (or possible) to obtain all 11 testing features for every system of interest that one wishes to test. Therefore, it might be more efficient to reduce a model's accuracy in favor of reducing its complexity and number of features. We now aim to identify the most predictive features of the most successful model identified above. As such, we implemented a Permutation Feature Importance algorithm, where we quantified the decrease in the models' precision as a result of shuffling the values of a single feature. This effectively disassociates the feature values from each sample's labels. The relative magnitude of the drop in the model's accuracy is proportional to the relative importance of the shuffled feature. This drop is then normalized with respect to all other shuffled features, and a feature importance ranking is created. In our case, we chose to shuffle and measure the score effect of each feature 50 times before continuing on to the next one.

Feature	Ranking	Rel. Importance	Marginal Accuracy
D/P	1	0.17	0.58
D/T	2	0.16	0.77
F_w	3	0.12	0.85
U_{nd}	4	0.07	0.88
PW_{nd}	5	0.07	0.92

TABLE III. Relative feature importance for the best-performing classifier. The rightmost column identifies the accuracy of a model trained and tested exclusively with the top "n" features.

Table III reports the top 5 features for the optimized randomized trees classifier, where the ratio between particle diameter and pore size (D/P), the ratio between the particle diameter and the geometric thickness (D/T), and the standardized particle flux (F_w) are the top three features. This is consistent with the experimental findings of Gerber et. al., 2018 [8] and Robert de Saint Vincent et. al., 2016 [3], which show that clogging is highly correlated to these properties. The 4th and 5th most predictive features are the non-dimensional fluid velocity (U_{nd}) and the non-dimensionalized attractive forces (PW_{nd}); this is also consistent with previous findings

[2, 7, 40].

More surprising, however, is the relatively small predictive power that D/P has by itself, where a model trained and tested with only said feature can only predict clogging/unclogging 0.58 percent of the time (just slightly better than a coin flip). The addition of the next top two features increases prediction accuracy to about 0.85 for both classifiers, meaning that the bulk of the simulated clogging cases can be explained/predicted by just 2 complementary physical processes: particle size exclusion and particle flux. The remaining cases rely on more complex physics, where it is important to consider viscous flow-particle couplings and particle-particle-wall attractions.

Contrastingly, porosity, pore size heterogeneity, and grain heterogeneity do not have much influence on clogging. The low dependence of clogging on porosity has been observed before by Wyss et. al., 2006 [12] and Zuriguel et. al., 2014 [4]. To the best of our knowledge, the apparent independence of clogging on spatial heterogeneity has not been studied or observed in any previous study.

E. Training and Identifying Best Regressors

Training and testing ML regressors followed the same procedure outlined in Section III C. The only difference was the fact that we are now aiming to predict the CN associated with each sample, as opposed to its discrete label (i.e. Clogged vs Unclogged). Given that the degree of clogging is a non-discrete latent variable, it can be argued that using regressors to predict a CN might actually be more useful than binary classification. However, it is important to consider that the CN is an abstract metric; it is not obvious how this number's magnitude reflects actual physical states. For this reason we make no value-judgement as to which approach might be more useful and instead present the results of both types of predictors.

The results of our training, testing, and optimization procedure are shown in Table IV, where we evaluate regressor performance by quantifying the R^2 and Mean Absolute Error (MAE) obtained from plotting a "Predicted log(CN)" vs "Measured log(CN)" curve (Figure 5). In this case, we used the log(CN) as our target feature in order to ensure we don't fit a model that is overly biased towards predicting high values of CN.

Table IV clearly shows that decision tree-based ML models are the best predictors for clogging mechanisms in porous media, closely followed by neural networks (MLPs). Both a description and the trained algorithm for the best performing optimized model (another Extremely Randomized Trees algorithm) can be found in the Supplementary Materials.

Regressor	R^2	MAE
Support-Vector Machine	0.54	0.65
Linear Regression	0.57	0.66
K-Nearest Neighbors	0.77	0.41
Multilayer Perceptron	0.89	0.31
Random Forests	0.91	0.21
Extremely Randomized Trees	**0.93**	**0.19**

TABLE IV. Clogging prediction performance of the chosen ML regressors. The section in bold represents the regressor with the best performance.

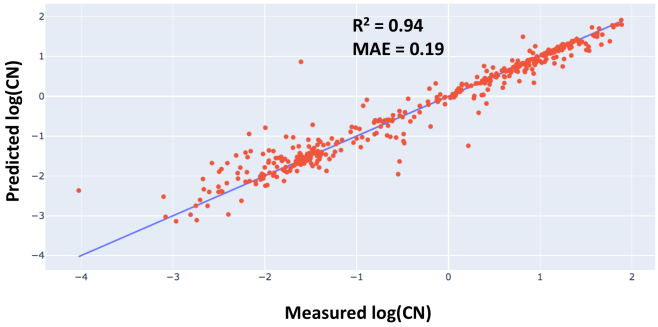

FIG. 5. Optimized Extremely Randomized Trees regression results.

F. Regression Feature Analysis

Just as we did for classifiers, we now turn to evaluate feature importance for our best regressor through a Permutation Feature Importance algorithm. The results shown in Table V confirm the feature ranking that we obtained from our best classifier, as we obtain the same 5 top predictive features as we did before. The top two features are once again related to size exclusion effects (D/P and D/T). This time however, the particle-wall attraction (PW_{nd}) is ranked as the third most predictive feature, followed by the standardized particle flux (F_w) and non-dimensional fluid velocity (U_{nd}). This difference in ranking is most likely due to differing implementations of the node-split criteria for both tree-based algorithms (shown the Supporting Information). However, the fact that both models obtained the same top 5 features using different approaches is a strong indication that these features are indeed significant.

Just as we did before, we observe that size exclusion is a necessary but not sufficient feature when it comes to predicting clogging in our simulations (a model trained with just the ratio between particle diameter and pore size (D/P and D/T) obtains a R^2 value of 0.56). Therefore, it is strictly necessary to at least consider the fluid flow rate, the number of particles injected, and attraction forces in order to create a model with a reasonable predictive power.

Feature	Ranking	Rel. Importance	Marginal R^2
D/P	1	0.27	0.53
D/T	2	0.20	0.57
PW_{nd}	3	0.13	0.62
F_w	4	0.11	0.67
U_{nd}	5	0.10	0.80

TABLE V. Tabulated relative feature importance for the top 5 Extremely Randomized Trees regression features. The rightmost column identifies the score metrics of the model if it is trained and tested exclusively with the top "n" features.

IV. PREDICTING CLOGGING IN REAL POROUS SYSTEMS

Finally, having trained, identified, and interpreted the best performing ML classifier and regressor, we now proceed to test their ability to predict clogging in additional numerical and experimental samples.

To do so we obtained the features outlined in Section III B from three different data sets produced by [13], [4], and [11]. These ranged from single-pore systems, to quasi 2-D ordered multi-pore systems, to 3-D complex porous media and covered multiple scales, flow rates, particle concentrations, and particle-particle-wall interactions. Please refer to the Supporting Information for a table containing all of the samples' testing features and labels.

The resulting predictions are consistent with the ones performed in Section III C. The trained classifier was able to predict clogging in 11 out of 12 systems when using all 11 features in a set with 6 clogged and 6 unclogged samples. The prediction power of the algorithm dropped slightly to 10 out of 12 when using only the top 5 features specified in Section III D.

Unfortunately, we could not test our trained regression algorithm against these additional samples, as it was impossible to calculate a CN for said systems (the aforementioned studies did not measure a particle velocity distribution and/or inter-particle distances).

V. CONCLUSIONS

In this study, we combined direct numerical simulations with machine learning approaches to predict clogging in heterogeneous porous media. To do so, we developed a computational workflow designed to simulate particle flow through randomly-generated porous media over thousands of different experimental and parametric conditions. One particular challenge that arose was the fact that clogging is a continuous process, not a discrete one. For this reason, we also developed the concept of a "Clogging Number" in order to properly characterize the level of clogging in our computational simulations.

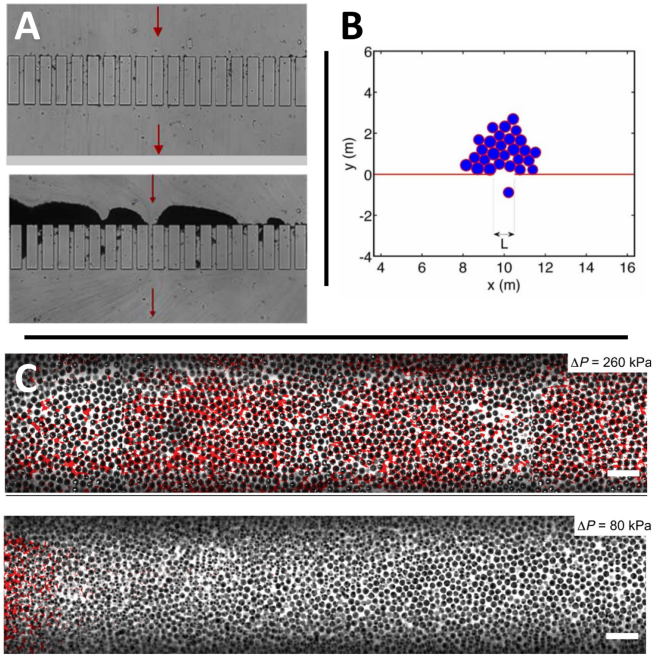

FIG. 6. Sample images of the experimental samples used to verify and test our algorithm. A) Experiments performed by [13] where microscopic polystyrene particles where injected at different concentrations and fluid flow rates into an ordered 2-D micromodel. B) Simulations of pedestrian evacuation of a crowded room performed by [4], where they varied the target flow speed and the size of the exit door. C) Experiments performed by [11] where they quantified the degree of clogging within a 3-D micromodel by varying the fluid flowrate and inter-particle forces.

Furthermore, in order to standardize our results we developed 11 non-dimensional training features that can be measured *a-priori* from any system involving the flow of solid particles through a static obstacle field.

After training, testing, and optimizing several classifiers and regressors we concluded that the best performing classifier was an Extremely Randomized Trees algorithm. This classifier was able to achieve values of 0.96, 0.93, and 0.99 for label accuracy, precision, and ROC, respectively. Similarly, the best regressor was a Extremely Randomized Trees algorithm that was able to achieve a R^2 value of 0.93 when used to predict the samples' Clogging Number. The interested reader can access all the associated data, hyperparameters, and trained algorithms in the Supplementary Materials.

Feature importance analysis of both machine learning approaches showed that the 5 most predictive features for clogging in porous media are: the ratio between particle diameter and the pore size, the flux of particles through the porous medium, particle-particle/wall attraction forces, and the fluid velocity. We also demonstrated that, although all these features are necessary to create an accurate prediction model, none of these are sufficient by themselves. Furthermore, we observed that

the system porosity and the spatial heterogeneity of the porous medium did not play a significant role in determining whether a system clogs or not.

In the last section of this study we tested the predictive power of our trained classifier in 12 experimental cases from three different data sets, showing a sustained prediction accuracy of 0.92 (11 out of 12) when used with all 11 engineered features and an accuracy of 0.83 (10 out of 12) when using only the top 5 most predictive features.

The result of this investigation is a generalized clogging prediction algorithm that can accurately predict clogging in porous media, a tool which we hope can help improve the design process of engineered porous systems. This is the first time that anyone has probed such a large parametric phase space and/or applied machine learning to try and describe this difficult problem, to the best of our knowledge. Additional work is required the evaluate the ability of our models to predict clogging in a broader range of conditions and physical systems. To do so we would need expand our simulated parameter space into different particle shapes, different grain geometries, and a larger range of fluid velocities and simulation times. We hope that this investigation spurns further work into integrating machine learning approaches with numerical simulations in order to probe and characterize stochastic physical phenomena.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, Division of Earth Sciences, Early Career program through Award EAR-1752982, and by the Mary and Randall Hack '69 Research Fund provided by the High Meadows Environmental Institute at Princeton University.

[1] I. L. Molnar, W. P. Johnson, J. I. Gerhard, C. S. Willson, and D. M. O'Carroll, Predicting colloid transport through saturated porous media: A critical review, Water Resources Research **51**, 6804 (2015).

[2] T. Phenrat, H. J. Kim, F. Fagerlund, T. Illangasekare, R. D. Tilton, and G. V. Lowry, Particle size distribution, concentration, and magnetic attraction affect transport of polymer-modified Fe0 nanoparticles in sand columns, Environmental Science and Technology **43**, 5079 (2009).

[3] M. Robert De Saint Vincent, M. Abkarian, and H. Tabuteau, Dynamics of colloid accumulation under flow over porous obstacles, Soft Matter **12**, 1041 (2016).

[4] I. Zuriguel, D. R. Parisi, R. C. Hidalgo, C. Lozano, A. Janda, P. A. Gago, J. P. Peralta, L. M. Ferrer, L. A. Pugnaloni, E. Clément, D. Maza, I. Pagonabarraga, and A. Garcimartín, Clogging transition of many-particle systems flowing through bottlenecks, Scientific Reports **4**, 1 (2014).

[5] B. Ding, C. Li, M. Zhang, F. Ji, and X. Dong, Effects of pore size distribution and coordination number on the prediction of filtration coefficients for straining from percolation theory, Chemical Engineering Science **127**, 40 (2015).

[6] M. Sahimi and A. O. Imdakm, Hydrodynamics of particulate motion in porous media, *Physical Review Letters*, Physical Review Letters **66**, 1169 (1991).

[7] D. Liu, P. R. Johnson, and M. Elimelech, Colloid Deposition Dynamics in Flow Through Porous Media: Role of Electrolyte Concentration, Environmental Science and Technology **29**, 2963 (1995).

[8] G. Gerber, S. Rodts, P. Aimedieu, P. Faure, and P. Coussot, Particle-Size-Exclusion Clogging Regimes in Porous Media, Physical Review Letters **120**, 10.1103/PhysRevLett.120.148001 (2018).

[9] M. Mirabolghasemi, M. Prodanovic, D. DiCarlo, and H. Ji, Prediction of empirical properties using direct pore-scale simulation of straining through 3D microtomography images of porous media, Journal of Hydrology **529**, 768 (2015).

[10] N. H. Pham and D. V. Papavassiliou, Effect of spatial distribution of porous matrix surface charge heterogeneity on nanoparticle attachment in a packed bed, Physics of Fluids **29**, 5492 (2017).

[11] N. Bizmark, J. Schneider, R. D. Priestley, and S. S. Datta, Multiscale dynamics of colloidal deposition and erosion in porous media, Science Advances **6**, 2530 (2020).

[12] H. M. Wyss, D. L. Blair, J. F. Morris, H. A. Stone, and D. A. Weitz, Mechanism for clogging of microchannels, Physical Review E 10.1103/PhysRevE.74.061402 (2006).

[13] G. C. Agbangla, É. Climent, and P. Bacchin, Experimental investigation of pore clogging by microparticles: Evidence for a critical flux density of particle yielding arches and deposits, Separation and Purification Technology **101**, 42 (2012).

[14] M. Auset and A. A. Keller, Pore-scale visualization of colloid straining and filtration in saturated porous media using micromodels, Water Resources Research **42**, 10.1029/2005WR004639 (2006).

[15] G. Gerber, M. Bensouda, D. A. Weitz, and P. Coussot, Self-Limited Accumulation of Colloids in Porous Media, Physical Review Letters **123**, 158005 (2019).

[16] X. Li, C. L. Lin, J. D. Miller, and W. P. Johnson, Role of grain-to-grain contacts on profiles of retained colloids in porous media in the presence of an energy barrier to deposition, Environmental Science and Technology **40**, 3769 (2006).

[17] X. Li, C. L. Lin, I. D. Miller, and W. P. Johnson, Pore-scale observation of microsphere deposition at grain-to-grain contacts over assemblage-scale porous media domains using x-ray microtomography, Environmental Science and Technology **40**, 3762 (2006).

[18] C. Chen, B. L. Lau, J. F. Gaillard, and A. I. Packman, Temporal evolution of pore geometry, fluid flow, and solute transport resulting from colloid deposition, Water Resources Research **45**, 10.1029/2008WR007252 (2009).

[19] D. C. Mays, O. T. Cannon, A. W. Kanold, K. J. Harris, T. C. Lei, and B. Gilbert, Static light scattering resolves

colloid structure in index-matched porous media, Journal of Colloid and Interface Science **363**, 418 (2011).
[20] J. Schneider, R. D. Priestley, and S. S. Datta, Using colloidal deposition to mobilize immiscible fluids from porous media, Physical Review Fluids **6**, 18 (2021).
[21] G. C. Agbangla, P. Bacchin, and E. Climent, Collective dynamics of flowing colloids during pore clogging, Soft Matter **10**, 6303 (2014).
[22] F. Civan, Non-isothermal Permeability Impairment by Fines Migration and Deposition in Porous Media including Dispersive Transport, Transport in Porous Media **85**, 233 (2010).
[23] M. R. Wiesner, M. C. Grant, and S. R. Hutchins, Reduced permeability in groundwater remediation systems: Role of mobilized colloids and injected chemicals, Environmental Science and Technology **30**, 3184 (1996).
[24] F. A. Weber, A. Voegelin, R. Kaegi, and R. Kretzschmar, Contaminant mobilization by metallic copper and metal sulphide colloids in flooded soil, Nature Geoscience **2**, 267 (2009).
[25] A. B. Kersting, D. W. Efurd, D. L. Finnegan, D. J. Rokop, D. K. Smith, and J. L. Thompson, Migration of plutonium in ground water at the Nevada Test Site, Nature **397**, 56 (1999).
[26] J. N. Ryan and M. Elimelech, Colloid mobilization and transport in groundwater, Colloids and Surfaces A: Physicochemical and Engineering Aspects **107**, 1 (1996).
[27] J. Sirignano, J. F. MacArt, and J. B. Freund, DPM: A deep learning PDE augmentation method with application to large-eddy simulation, Journal of Computational Physics **423**, 267 (2020).
[28] J. Kocić, N. Jovičić, and V. Drndarević, An end-to-end deep neural network for autonomous driving designed for embedded automotive platforms, Sensors (Switzerland) **19**, 267 (2019).
[29] F. Messina, T. Tosco, and R. Sethi, On the failure of upscaling the single-collector efficiency to the transport of colloids in an array of collectors, Water Resources Research **52**, 5492 (2016).
[30] G. Boccardo, E. Crevacore, R. Sethi, and M. Icardi, A robust upscaling of the effective particle deposition rate in porous media, Journal of Contaminant Hydrology **212**, 3 (2018), arXiv:1702.04527.
[31] M. Hilpert and W. P. Johnson, A Binomial Modeling Approach for Upscaling Colloid Transport Under Unfavorable Attachment Conditions: Emergent Prediction of Nonmonotonic Retention Profiles, Water Resources Research **54**, 46 (2018).
[32] R. Jäger, M. Mendoza, and H. J. Herrmann, Channelization in porous media driven by erosion and deposition, Physical Review E **95**, 5492 (2017), arXiv:1609.03746.
[33] J. Zhao and T. Shan, Coupled CFD-DEM simulation of fluid-particle interaction in geomechanics, Powder Technology **239**, 248 (2013).
[34] S. Remond, DEM simulation of small particles clogging in the packing of large beads, Physica A: Statistical Mechanics and its Applications **389**, 4485 (2010).
[35] S. Natsui, S. Ueda, H. Nogami, J. Kano, R. Inoue, and T. Ariyama, Gas-solid flow simulation of fines clogging a packed bed using DEM-CFD, Chemical Engineering Science **71**, 274 (2012).
[36] T. Li, S. Schlüter, M. I. Dragila, and D. Wildenschild, An improved method for estimating capillary pressure from 3d microtomography images and its application to the study of disconnected nonwetting phase, Advances in Water Resources **114**, 249 (2018).
[37] E. Barthel, Adhesive elastic contacts: JKR and more, Journal of Physics D: Applied **41**, 20 (2008).
[38] Z. Y. Zhou, S. B. Kuang, K. W. Chu, and A. B. Yu, Discrete particle simulation of particle-fluid flow: Model formulations and their applicability, Journal of Fluid Mechanics **661**, 482 (2010).
[39] P. Geurts, D. Ernst, and L. Wehenkel, Extremely randomized trees, Machine Learning **63**, 3 (2006).
[40] S. Aachanta, Nonequlibirium swelling and cappillary pressure relations for colloidal systems., Journal of Colloid and Interface Science , 266 (1994).