高分子マテリアルズ・インフォマティクス（MI）の確立に向けた記述子の開発

課題責任者
鷲津　仁志　　兵庫県立大学大学院情報科学研究科　京都大学 ESICB

著者
清水　陽平　　兵庫県立大学大学院シミュレーション学研究科

要旨
高分子マテリアルズ・インフォマティクス（MI）の実現にあたっては，材料特性に影響する高分子特有の高次構造を適切な条件でシミュレーションすることと，機械学習可能な記述子を見出すことが必要であるが，未だ完全に実現された例はない．本研究では，大規模 MD と機械学習の組み合わせにより高分子 MI の実現可能性の模索を主目的としている．アプローチとして，膨大な組み合わせとなる探索対象の問題を解決するため，高次構造を抽象化した記述子により特性の方向性を指し示す AI の実現を検討している．高次構造を抽象化する手法として，位相幾何学のパーシステントホモロジーに着目した．第一段階として，ポリマーメルトを対象として全原子 MD 計算を行い，その座標をパーシステントホモロジーによって解析することで高次構造の表現が可能であることを示した．第二段階として，様々な種類のポリマーメルトについて，同様の解析を行い，記述子としてランダムフォレストによる機械学習を行うことで，従来のフィンガープリントによる記述子と比べ，高い予測精度を得ることができた．

キーワード
Materials Informatics, Higher-order structure, Dielectric constant, Persistent homology, Machine learning

1　パーシステントホモロジーによる高次構造の表現

1.1　全原子 MD におけるポリマーメルトの高次構造変化

MD の周期境界条件においては，一般にセルサイズの物性への依存はないものと考えられるが，ポリマーについては，依存性についての報告例がいくつかある [1, 2]．同じ密度（0.98 に設定）でセルサイズやポリマー鎖長が異なる NBR(Acrylonitrile-Butadiene) ポリマーメルトの全原子 MD において（図 1a-c，表 1），電場を与え，平衡時の比誘電率を計算すると，明らかな依存性が生じた．このとき，動径分布関数 (RDF) は変化がないことから，隣接原子に近い範囲ではなく，より高次な構造が物性に寄与していることが示唆された（図 2a, b）．計算には LAMMPS(Version Lammps-22Aug18, http://lammps.sandia.gov) を用い，力場は Dreiding を使用した．電荷は AM1-BCC により得た．NPT アンサンブル 500K，1000atm，1ns 計算を行ったのち，NVT アンサンブル 300K，1ns でアニーリングを行い，初期構造を得た．得られた初期構造に電場 1.0×10^{-3} V/nm を 1 方向に印加し，1ns 後に得られた平衡構造を評価に用いた．比誘電率は式 (1) により計算した [3].

$$\epsilon_r = \frac{P}{\epsilon_0 E} + 1 \qquad (1)$$

ここで，ϵ_r は比誘電率, ϵ_0 は真空の誘電率, E は電場の大きさ, P は分極である．

1.2　パーシステントホモロジーを用いたポリマーメルト高次構造の解析

そこで，近年着目されているパーシステントホモロジーを適用することで，高次構造を表現できるか検証を行った．パーシステントホモロジーは位相幾何学（トポロジー）の分野であり，MD の全原子の 3 次元座標情報から 2 次元のパーシステント図を求める．ここではアルファ複体とアルファフィルタレーションを考える．原子の座標から仮想的に円を大きくしていき，円によって囲まれた領域が発生した大きさを birth time，さらに円を大きくしていき，円によって囲まれた領域が消滅した大きさを death time と定義する．横軸を birth time，縦軸を death time とし，すべての領域についてその関係をプロットする．これをパーシステント図という [4, 5]．3 次元座標では 0 次から 2 次までのパーシステント図が存在し，2 次のパーシステント図は空隙の大きさを表す (図 1d)．セルサイズやポリマー鎖長による MD の全原子座標からそれぞれ 2 次のパーシステント図を求めた結果，動径分布関数と異なり，明確な空隙構造の変化が認められた．セルとポリマーが接触しやすい小さなセルの条件では，パーシステント図が凝集し，大きなセルでは拡散した．さらに，定量的に比較するため，パーシステントベッチ数 (PBNs)[6, 7] を求めると，セルサイズが大きくなるにつれ，グラフが平滑となることが確認できた（図 2c）．これらの実験的事実から，

ポリマーメルトの高次構造を表現する手法としてパーシステント図が適していることが分かった [8]．なお，パーシステント図の計算は Homcloud(Version 2.8.1, http://www.wpi-aimr.tohoku.ac.jp/hiraoka_lab/homcloud-english.html) を用いた [9, 10]．

2 パーシステントホモロジーを記述子としたポリマーメルトの比誘電率予測

2.1 記述子による比誘電率の予測精度の比較

パーシステント図はベクトル化により決まったビット長で表現できることから機械学習における記述子として利用できる [11]．そこで，モノマー比率やブレンドを行ったポリマーメルトの教師データ 229 個の全原子 MD を行い，300K における原子座標と比誘電率の結果を得た．目的変数を比誘電率とし，説明変数を原子座標から得た 2 次のパーシステント図をベクトル化したものと，比較として一般的な固定長バイナリによる Fingerprints 表現形式である ECFP6 の 2 種類として，学習と予測精度の確認を行った．学習にあたっては，教師データの数より説明変数が多い，いわゆるスパース問題となっていることから，変数選択手法としてランダムフォレストを繰り返し実施することで，多く選択されたベクトルを記述子とした．選択した記述子を用い，再度ランダムフォレストモデルによる回帰予測モデルを構築した．これらの学習には Scikit-learn(Version 0.23.0, https://scikit-learn.org/) を用いた．その結果，高次構造を表現できない ECFP6 と比較して，パーシステント図の記述子は良好な予測精度を得られることを確認した（図 3）．

2.2 抽象化した記述子による探索空間の定義

上記教師データはある一定の条件下において網羅的な条件で取得したものの，実際のポリマーは無限の探索空間が考えられ，通常の機械学習では予測モデルの構築が困難である．しかし，パーシステント図においては，原子の情報や高次構造が抽象化されて表現されており，具体的な材料の組成を指し示すことは難しいが，パラメータをどの方向に持っていけばよいか示唆できることが期待される．実際の材料開発における材料研究者のニーズも，具体的に完成された組成条件予測よりも，次に行う実験の指針を与えてくれる AI がまずは求められると考えられる．パーシステント図による記述子の空間を一定の範囲に定義できれば，対応する目的変数である比誘電率がどのように分布するか予測できる．そこで，パーシステント図によるベクトル化された記述子を自己組織化マップ (SOM) により 2 次元座標変換し，対応する比誘電率をラベル表記した．その結果，比誘電率の傾向は SOM 上で滑らかに分布することが分かった（図 4a）．ガウス過程回帰を組み合わせることで，目的の比誘電率に対応する未知の高次構造を予測可能であることが分かった（図 4b）．

3 謝辞

本研究の成果は，海洋研究開発機構の地球シミュレータ ES3 を用いて得られたものです（課題 ID:G6192）．
本研究は科研費（JP18K18813, JP19H05718, JP20H02058）の助成を受けたものです．

参考文献

[1] Toshiki Mima, Tetsu Narumi, Shun Kameoka, and Kenji Yasuoka. Cell size dependence of orientational order of uniaxial liquid crystals in flat slit. *Molecular Simulation*, Vol. 34, No. 8, pp. 761–773, 2008.

[2] Sezen Curgul, Krystyn J. Van Vliet, and Gregory C. Rutledge. Molecular dynamics simulation of size-dependent structural and thermal properties of polymer nanofibers. *Macromolecules*, Vol. 40, No. 23, pp. 8483–8489, 2007.

[3] Hu Taotao. The predicted dielectric constant of an amorphous pvdf changing with temperature by molecular dynamics simulations. *Int. J. of Electrochem. Sci.*, Vol. 13, pp. 10088–10100, 11 2018.

[4] H. Edelsbrunner and J. Harer. Persistent homology - a survey. *Ser. Contemp. Sppl. Math.*, Vol. 453, , 2008.

[5] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, Vol. 33, pp. 249–274, 2005.

[6] Zhenyu Meng, D. Vijay Anand, Yunpeng Lu, Jie Wu, and Kelin Xia. Weighted persistent homology for biomolecular data analysis. *Scientific Reports*, Vol. 10, No. 1, p. 2079, Feb 2020.

[7] D. Vijay Anand, Zhenyu Meng, Kelin Xia, and Yuguang Mu. Weighted persistent homology for osmolyte molecular aggregation and hydrogen-bonding network analysis. *Scientific Reports*, Vol. 10, No. 1, p. 9685, Jun 2020.

[8] Yohei Shimizu, Takanori Kurokawa, Hirokazu Arai, and Hitoshi Washizu. Higher-order structure of polymer melt described by persistent homology. *Scientific Reports*, Vol. 11, No. 1, p. 2274, Jan 2021.

[9] Marcio Gameiro, Yasuaki Hiraoka, and Ippei Obayashi. Continuation of point clouds via persistence diagrams. *Physica D: Nonlinear Phenomena*, Vol. 334, pp. 118–132, 2016.

[10] Ippei Obayashi. Volume-optimal cycle: Tightest representative cycle of a generator in persistent homology. *SIAM J. Appl. Algebra Geometry*, Vol. 2, No. 4, pp. 508–534, 2018.

[11] Ippei Obayashi, Yasuaki Hiraoka, and Masao Kimura. Persistence diagrams with linear machine learning models. *J Appl. and Comput. Topology*, Vol. 1, No. 3, pp. 421–449, 2018.

表 1: 高次構造検証のためのポリマーメルト組成と平衡温度条件

サンプル名	初期セルサイズ (nm)	Acrlonitrile 比率 (mol%)	Butadiene 比率 (mol%)	ポリマー鎖長 (mer)	平衡温度 (K)
Sample 1	2.0038	95	5	10, 20, 40	150 ,300, 450

表 2: 機械学習に用いたポリマーメルトの組成条件

ポリマー種類	モノマー組み合わせ（5mol% 刻みで設定）	サンプル数
NBR	Acrlonitrile, Butadiene	19
ACM	Ethyl acrylate, Butyl acrylate, Methoxyethyl Acrylate	64
FKM	Vinylidene floride, Hexafluoropropylene, Tetrafluoroethylene	45
VMQ	Dimethylsiloxane, Methylvinylsiloxane, Methlphenylsiloxane	19
ブレンドポリマー	上記ポリマーをランダムにブレンド	82
	Total	229

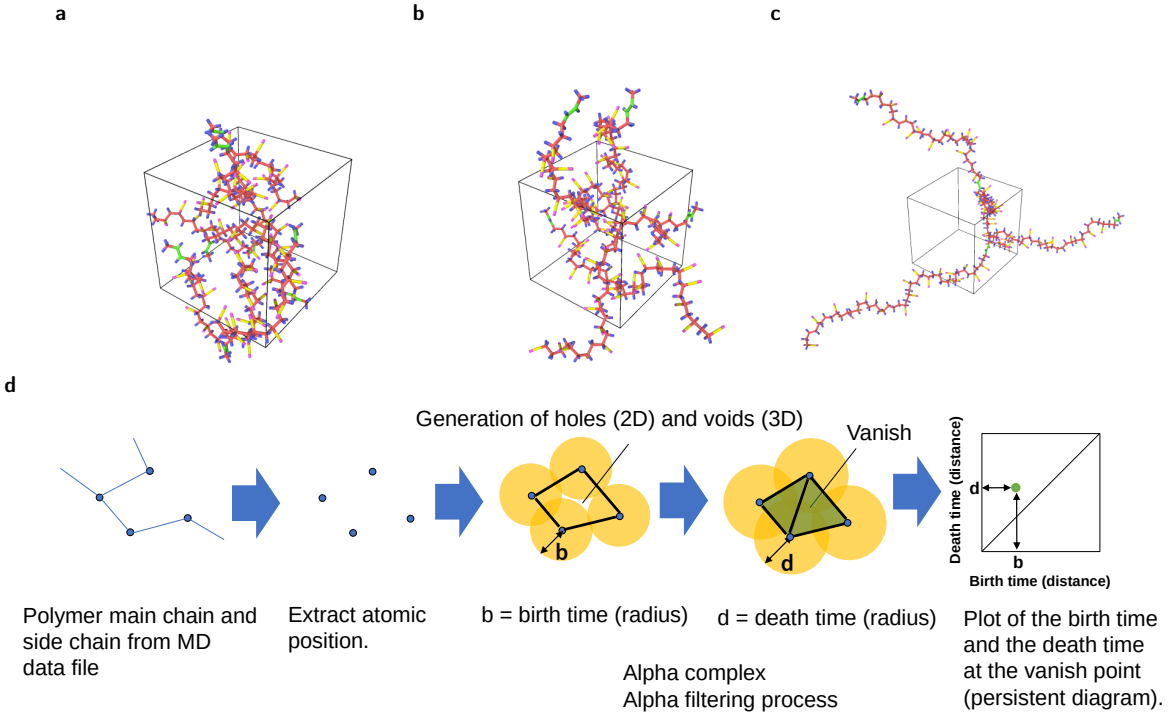

図 1: Sample 1 におけるポリマーメルトの鎖長とセルサイズのイメージ．(a) 10mer，(b) 20mer，(c) 40mer．(b), (c) ではセルからポリマー鎖が飛び出している．(d) 全原子 MD の原子座標を用いたパーシステント図の計算方法.

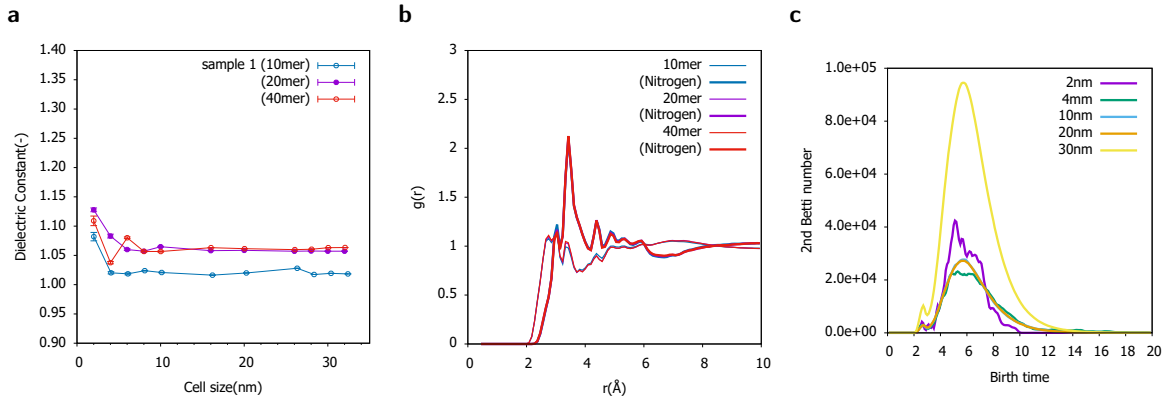

図 2: (a) Sample 1，300K におけるセルサイズとポリマー鎖長による比誘電率の変化．エラーバーは $\pm 3\sigma$ を示す．(b) Sample 1，300K のセルサイズ 20nm における動径分布関数 (RDF) の変化．(Nitrogen) は窒素原子周りに着目した場合の計算結果．(c) Sample 1，300K，10mer の 2 次のパーシステント図より求めたパーシステントベッチ数 (PBNs) の変化．セルサイズが異なり原子数が等しくないが，原子数を合わせる処理を行っても傾向は変わらない．

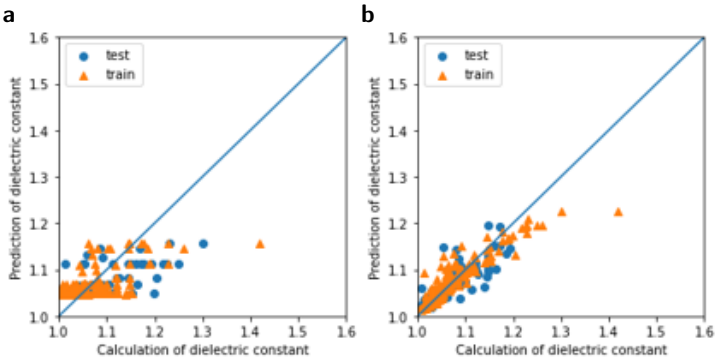

図 3: 比誘電率の MD 計算結果と機械学習による予測結果．(a) フィンガープリント (ECFP6) を記述子として用いた場合． 予測精度 Test/Train = 0.37/0.39．(b) ベクトル化したパーシステント図からランダムフォレストによる変数選択を行ったものを記述子として用いた場合．予測精度 Test/Train = 0.58/0.81．

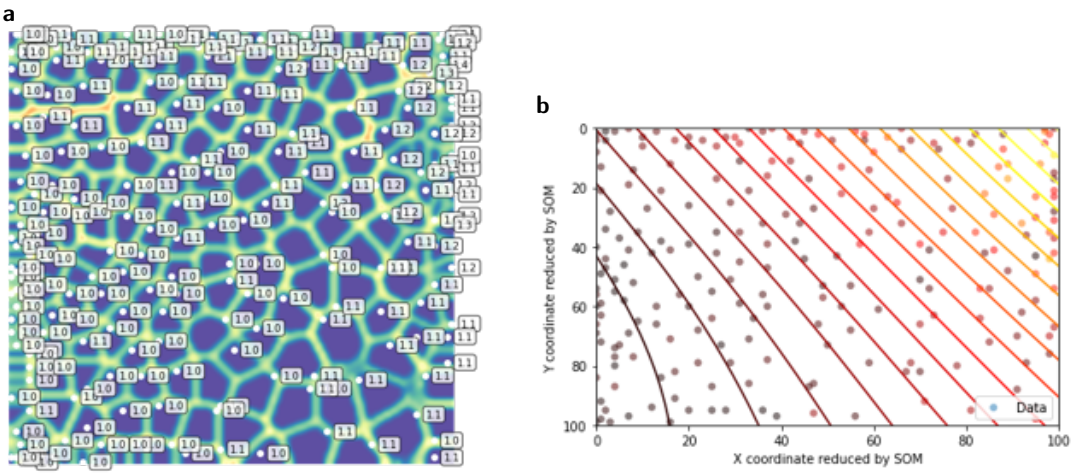

図 4: (a) 自己組織化マップ (SOM) によるポリマーメルト探索空間の分布．ラベルの数字は各ポリマーの比誘電率を表す． (b) ガウス過程回帰による比誘電率の分布．計算していない空間も比誘電率を幅を持たせて予測できる．

Development of Descriptors for Establishing Polymer Materials Informatics (MI)

Project Representative

Hitoshi Washizu, Graduate School of Information Studies, University of Hyogo,

Elements Strategy Initiative for Catalysts and Batteries (ESICB), Kyoto University

Author

Yohei Shimizu, Graduate School of Simulation Studies, University of Hyogo

Abstract

In order to realize polymer materials informatics (MI), it is necessary to simulate high-order structures peculiar to polymers that affect material properties under appropriate conditions and to find machine-learnable descriptors. , There is no example that has been completely realized yet. The main purpose of this study is to explore the feasibility of polymer MI by combining large-scale MD and machine learning. As an approach, in order to solve a huge number of combinations of search target problems, we are studying the realization of AI that indicates the direction of characteristics by using a descriptor that abstracts the higher-order structure. As a method of abstracting higher-order structures, we focused on persistent homology of topology. As the first step, it was shown that the higher-order structure can be expressed by performing all-atom MD calculation for the polymer melt and analyzing its coordinates by persistent homology. As the second step, by performing the same analysis for various types of polymer melts and performing machine learning using a random forest as the descriptor, higher prediction accuracy can be obtained compared to the descriptor by the conventional fingerprint.

Keywords

Materials Informatics, Higher-order structure, Dielectric constant, Persistent homology, Machine learning

1 Representation of higher-order structures by persistent homology

Under the periodic boundary conditions of MD, it is generally considered that there is no dependence on the physical properties of cell size, but there are some reports on the dependence on polymers.[1]. In all-atom MD of NBR (Acrylonitrile-Butadiene) polymer melt with the same density (set to 0.98) but different cell size and polymer chain length, When an electric field is applied and the dielectric constant at equilibrium is calculated, a clear dependence arose. At this time, the radial distribution function (RDF) does not change, suggesting that higher-order structures, rather than those close to adjacent atoms, contribute to the physical characteristics. (Fig. 1a, b) . We use LAMMPS(Version Lammps-22Aug18, http://lammps.sandia.gov) and Dreiding force. Charge has assigned at AM1-BCC. After calculating the NPT ensemble 500K, 1000atm and 1ns, annealing is performed with the NVT ensemble 300K and 1ns to obtain the initial structure.

Therefore, we have verified whether higher-order structures can be expressed by applying persistent homology, which has been attracting attention in recent years. Persistent homology is a field of topology,

and a two-dimensional persistent diagram is obtained from the three-dimensional coordinate information of all atoms of MD. Here, we consider alpha complex and alpha filtration. The circle is virtually enlarged from the coordinates of the atom, the size at which the area surrounded by the circle is generated is the birth time, and the circle is further increased, and the size at which the area surrounded by the circle disappears is death. Defined as time. The horizontal axis is birth time and the vertical axis is death time, and the relationship is plotted for all regions. This is called a persistent diagram [2]. In 3D coordinates, there are 0th to 2nd order persistent diagrams, and the 2nd order persistent diagrams represent the size of the voids. As a result of obtaining second-order persistent diagrams from all atomic coordinates of MD according to cell size and polymer chain length, a clear change in void structure was observed, unlike the radial distribution function. Under the condition of a small cell where the cell and the polymer easily contact, the persistent figure aggregated and diffused in the large cell. Furthermore, for quantitative comparison, the number of Persistent Betti numbers (PBNs) [3] is calculated, and it is confirmed that the graph became smoother as the cell size increased (Fig. 1 c). From these experimental facts, it is found that the persistent diagram is suitable as a method for expressing the higher-order structure of the polymer melt. The calculation of the persistent diagram is using Homcloud (Version 2.8.1, http://www.wpi-aimr.tohoku.ac.jp/hiraoka_lab/homcloud-english.html)[4].

2 Dielectric constant prediction of polymer melt using persistent homology as a descriptor

Since the persistent diagram can be represented by a bit length determined by vectorization, it can be used as a descriptor in machine learning [5]. Therefore, we have performed all-atom MD of 229 teacher data of polymer melt with monomer ratio and blending, and obtained the results of atomic coordinates and dielectric constant at 300K. Learning as two types, a vectorized quadratic persistent diagram obtained with the objective variable as the dielectric constant. ECFP6 which is a general fixed-length binary Fingerprints representation format is used for comparison. In learning, there are more explanatory variables than the number of teacher data. Therefore, by repeatedly executing a random forest as a variable selection method, a vector selected in large numbers iss used as a descriptor. Using the selected descriptor, a regression prediction model based on a random forest model is constructed again. Scikit-learn (Version 0.23.0, https://scikit-learn.org/) is used for these studies. As a result, it is confirmed that the descriptor of the persistent diagram can obtain better prediction accuracy than ECFP6, which cannot express the higher-order structure (Fig. 2).

3 Acknowledgements

This research used the computational resources of the Earth Simulator provided by the JAMSTEC (Project ID:G6192).
This study was supported by JSPS KAKENHI Grant Numbers JP18K18813, JP19H05718, and JP20H02058.

References

[1] Toshiki Mima, Tetsu Narumi, Shun Kameoka, and Kenji Yasuoka. Cell size dependence of orientational order of uniaxial liquid crystals in flat slit. *Molecular Simulation*, 34(8):761–773, 2008.

[2] H. Edelsbrunner and J. Harer. Persistent homology - a survey. *Ser. Contemp. Sppl. Math.*, 453, 2008.

[3] Zhenyu Meng, D. Vijay Anand, Yunpeng Lu, Jie Wu, and Kelin Xia. Weighted persistent homology for biomolecular data analysis. *Scientific Reports*, 10(1):2079, Feb 2020.

[4] Marcio Gameiro, Yasuaki Hiraoka, and Ippei Obayashi. Continuation of point clouds via persistence diagrams. *Physica D: Nonlinear Phenomena*, 334:118–132, 2016.

[5] Ippei Obayashi, Yasuaki Hiraoka, and Masao Kimura. Persistence diagrams with linear machine learning models. *J Appl. and Comput. Topology*, 1(3):421–449, 2018.

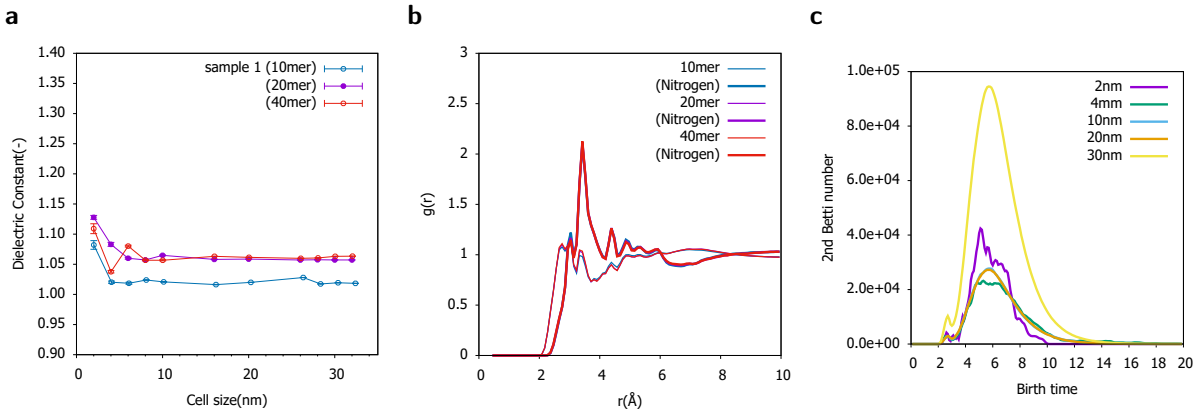

Figure 1: The change in dielectric constant (DC), the radial distribution function (RDF), Persistent Betti Number (PBNs). Error bars in DC represent $\pm 3\sigma$. (a) DC effect of polymer chain length in sample 1. (b) RDF effect of polymer chain length in sample 1 and 20nm cell size. Nitrogen means RDF around nitrogen atom. (c) 2nd PBNs of 300K and 10 mer.

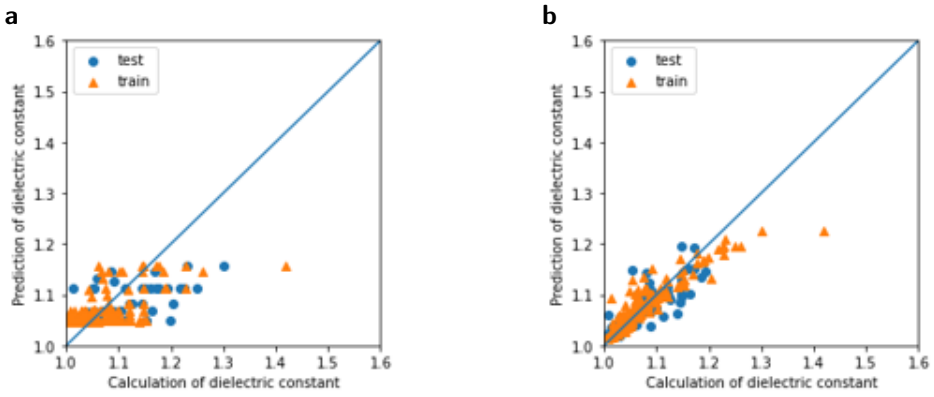

Figure 2: Calculation and prediction results of dielectric constant of melt polymer. (a) Fingerprints (ECFP6) are used as descriptor. Accuracy Test/Train = 0.37/0.39. (b) Selected vector from persistent diagram by random forest are used as descriptor. Accuracy Test/Train = 0.58/0.81.