ADVANCED FUNCTIONAL MATERIALS

Supporting Information

for *Adv. Funct. Mater.*, DOI: 10.1002/adfm.201301744

Stalking the Materials Genome: A Data-Driven Approach to the Virtual Design of Nanostructured Polymers

Curt M. Breneman, L. Catherine Brinson,* Linda S. Schadler,* Bharath Natarajan, Michael Krein, Ke Wu, Lisa Morkowchuk, Yang Li, Hua Deng, and Hongyi Xu*

DOI: 10.1002/adma.201301744

Stalking the Materials Genome: A Data-Driven Approach to the Virtual Design of Nanostructured Polymers

By *C. M. Breneman*, L.C. Brinson*, L.S. Schadler*, B. Natarajan, M. Krein, K. Wu, L. Morkowchuk, Y. Li, H. Deng, H. Xu*

Supporting Information

Informatics:

Polymer Representation: Molecular Operating Environment (MOE) version 2009 [1] was used to create 20-mer representations of polymers from monomer units using a modified version of MOE's built-in polymerization function. Terminal attachment points were capped with hydrogen. The backbone dihedral angles were set to 0° before geometry optimization. The highly flexible nature of the olignomer structures make the full exploration of conformation space impossible. To overcome this problem and find stable representation of polymer structure with sufficient 3D information, a slight tethering (10/300) was applied on the backbone atoms. Energy minimization was then performed with a MMFF94x forcefield with RMS tolerance of 0.05 kcal/mol. Alternate options for polymer length were explored,

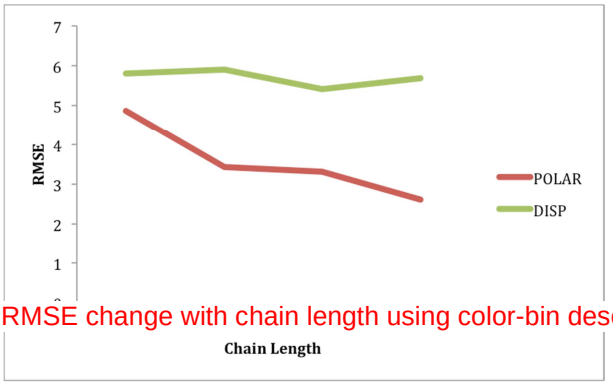

Fig 1. RMSE change with chain length using color-bin descriptor.

but a 20-mer length provided the best balance between computational time required while minimizing effects from capping hydrogens. However, 50-mers were created without optimization; they were necessary to minimize the influence of the end cap hydrogen atoms when calculating RECON/TAE descriptors, which is conformation-insensitive. Conformation search test was then performed to make sure the prediction uncertainty due to conformation variation is smaller than 5%.

Filler Representation: MOE version 2009 [1] was used to create a hollow, spherical silica shell with formula $Si_{30}O_{75}H_{30}$. Each silicon atom has an outward-facing –OH group; half of these were randomly assigned to be functional group attachment points in order to achieve a functionalization density of approximately 1/nm^2. In order to eliminate apparent molecular surface area in the interior of the shell, a small ($Si_4O_8H_8$) component was added to the center of the sphere for calculation of surface properties. Functional groups were added using MOE's QuaSAR-CombiGen tool, and completed beads were energy minimized using a MMFF94x forcefield with RMS tolerance of 0.01 kcal/mol.

Descriptors: Models for both the polar and dispersive components of polymer surface tension utilized full set of normalized RECON/TAE descriptor (those insensitive to the scale of chain length) and binned surface descriptors. Briefly, the normalized RECON/TAE descriptors describe the properties like local ionization potential, electrostatic potential and other eletron density distribution topology information, as a subset of Transferable Atomic Equivalent (TAE) fragments.[2]

A new set of surface descriptors was developed for this project using Connolly surfaces encoded with Electrostatic Potential (EP) or Active Lone Pair (ALP) mappings, as calculated by MOE.[3] The EP map covers a rage of -35 to 35 kcal/mole using an Ewald-type screened

molecular electrostatic potential.[4] The ALP surface map displays hydrophobic, mildly polar, and hydrogen-bonding regions. Resulting descriptors take the form of 8-bin probability density histograms for the intensity of each color channel (RGB) of each molecular surface (one histogram per color per surface), computed using R statistical software.[5] In addition, a grayscale histogram was created for each surface by the traditional NTSC color-to-luminance mapping of ((0.30*red) +(0.59*green) + (0.11*blue)). For polymer model, in order to further increase the stability of the descriptor values, a set of Gaussian smearing functions was applied so that each bin value was represented as a Gaussian function with the intergration value equal to origin value, with standard deviation to be one bin width. The smearing function was found to effective prevent descriptor value change due to the "value drifting" between two adjacent bins. A small value cut-off (1e-4) was also used to prevent uncertainty of extreme small value, which is a pure numerical problem.

Models for the polar component of filler surface tension utilized these color binned surface descriptors (excluding grayscale) using 256 histograms bins instead of 8; models were built using 8, 16, 32, 64, 128, and 256 histogram bins, and 256 yielded the best performance. Models for the dispersive component of filler surface tension utilized MOE's i3D descriptor package,[1] less those descriptors relying on AM1, PM3, and MNDO calculations.This descriptor setis conformationally sensitive and alignment free, containing energy, charge, and shape descriptors such as water accessible surface area, total potential energy, electrostatic potential, and dipole moment, for a total of 117 descriptors before feature selection.

Model Building

Preprocessing: Before machine learning algorithms were employed any descriptors showing zero variance across polymers were removed from the descriptor matrix. Remaining descriptors are standardized by subtraction of their mean and division by their standard

deviation across polymer samples. One descriptor was removed from any pair of descriptors showing a correlation coefficient greater than 0.85. Any single descriptor value lying outside [-6,6] after standardization was either changed to either -6 or 6 as appropriate (polymer dispersive model) or the descriptor was discarded (polymer polar model).Both filler models used a threshold of [-4,4], and values outside this range were changed to -4 or 4 as appropriate.

Feature Selection: Partial least squares (PLS) sensitivity feature selection was utilized for all MQSPR models. Briefly, descriptors are taken one at a time and varied uniformly first up and then down one standard deviation. A PLS model is built using the modified descriptor, and the resulting change in the RMSE from a non-modified descriptor matrix is recorded. Descriptors with high influence on RMSE are kept, and those with low influence are discarded. For the dispersive polymer model, four rounds of feature selection were completed where the least influential 30% of descriptors were discarded each round. For the polar polymer model, five rounds were completed where 20% of descriptors were discarded per round.Filler models underwent 12 rounds of 20% and ten rounds of 10% for polar and dispersive models, respectively. The number of rounds to complete and the percent of descriptors to discard were optimized using a grid search.

Descriptors remaining in the polymer dispersive model after feature selection are:

FDKNA10
8EP_Red
3EP_Gray
Ffuk5
FDGNA6
PIPMin
Flapl2
6ALP_Blue
7EP_Blue

Descriptors remaining in the polymer polar model after feature selection are:

FEP9
6EP_Green
PIPMax
7EP_Red
FDRNA10
6EP_Red
7EP_Blue
FSIKA3
6ALP_Green
7EP_Gray
6EP_Blue
2EP_Green

Descriptors remaining in the filler dispersive model after feature selection are:

FASA H
std_dim1
vsurf_Wp3
vsurf_CP
vsurf_Wp6
npr2
E
FASA-
Vsurf_CW1
Vsurf_IW8
E_ang
Vsurf_DW13
Vsurf_Wp7
E_oop
Vsurf_CW2

Descriptors remaining in the filler polar model after feature selection are:

124ALP_Red
180ALP_Red
205ALP_Red
212ALP_Red
86ALP_Green
16ALP_Blue
72EP_Red
96EP_Red
163EP_Red
19EP_Green
56EP_Green
75EP_Green
205EP_Blue
221EP_Blue

For color descriptors, the leading number indicates the histogram bin, EP or ALP designate the surface mapping, and the color indicates which color channel histogram is

referenced. Names for i3D descriptors are taken directly from MOE. Complete descriptions of these descriptors are available in MOE documentation.[1]

Machine Learning: Final models were all built on feature selected descriptor matrices. Classic multiple linear regression (MLR) models were built using the ln function in R.[5] Partial least squares (PLS) regression models were built using the pls package in R.[5] The number of latent variables to use was selected using 20 rounds of four-fold cross-validation. Epsilon-insensitive support vector machine (SVM) regression models were built on feature-selected descriptor matrices using R package e1071, a part of the R libsvm package.[6,7] Optimization of epsilon is handled internal to the e1071 package. Grid search hyperparameter selection for kernel parameter gamma and cost parameter C was completed using ten rounds of ten-fold bootstrap cross-validation, where the optimization criterion was cross-validated PRESS R^2. Gamma was searched over [1/50, 1/100, 1/200, 1/500, 1/1000, 1/(0.75*number_of_samples), and 1/number_of_descriptors]. C was searched over 1,2,3,4,5. Models using MLR, PLS, or SVM were chosen based on validated performance RMSE. Both polymer models use SVM, filler models use MLR and PLS for dispersive and polar components of surface tension, respectively. Bootstrapping was achieved by dividing experimental samples into random "training" and "validation" subsets. Models built using the training cases are used to predict the values of the validation cases. This process is repeated 100 times resulting in 100 "bootstrap" models, each trained on part of the original data. Each bootstrap model is then used to predict the values of all cases in each validation subset, so that the ultimate prediction result is the average of a large set of predictions made on each molecule in the original dataset, but only when it was part of the validation set, and not in the training set.Final bootstrap models were aggregated and predictions from the aggregate used for further studies.

Experiment:

Materials: PMMA (Sigma Alrich, M_w= 100000 g/mol), PS (Sigma Aldrich, M_w= 230000 g/mol) and PEMA (Scientific Polymer, M_w= 200000 g/mol) were used as procured. Prior to use, P2VP (Scientific Polymer, M_w= 200000 g/mol) was left overnight in a vacuum oven at 100ºC to evaporate residual monomer molecules. The nanoparticles (14±4nm Silica in MEK) were supplied by Nissan Inc. Octyldimethylmethoxysilane (CH_3-$(CH_2)_7$-$Si(CH_3)_2$-O-C_2H_5), Chloropropyldimethylmethoxysilane (Cl-C_3H_6-$Si(CH_3)_2$-O-C_2H_5) and Aminopropyldimethylethoxysilane (NH_2-C_3H_6-$Si(CH_3)_2$-O-C_2H_5) were procured from Gelest Inc. and used as received. These are monofunctional silanes, that ensure that only a single molecular layer is attached to the nanoparticle surface. HPLC grade anhydrous THF, anhydrous Toluene and Hexanes were procured from Fisher Scientific, Sigma Alrich and Mallincrodt Chemicals respectively. Silicon wafers, 2 inches in diameter, with a 200nm thick silica layer (Standard Deviation 2.58%) were procured from Silicon Quest Inc. High purity Formamide (>99.5%) and Diiodomethane (99%), the probe liquids used in surface energy measurement, were purchased from Sigma-Aldrich.

Methods:

Wafer surface modification: The 2 inch was broken into 4 even quarters, which were cleaned and hydroxylated by immersing in a Piranha Etch (70:30 solution of H_2SO_4 and 30% H_2O_2) at 80^0C for 12 hours. The hydroxylated pieces were rinsed in DI water, blown dry and left in a vacuum oven for 1-2 hours at 120^0C. The wafers were then surface modified through direct nucleophilic substitution of the silanol groups, by immersing in a 3% solution of the required monofunctional silane in Toluene for 3 days, in an inert atmosphere. These were then rinsed in ethanol, toluene and DI water, blown dry and vacuum dried at 120^0C for 10 minutes.

Particle surface modification: A 16 mL solution of colloidal silica particles (30 weight% in MEK) was diluted with 50 mL of THF. Around 0.5 mL of the silane was added and the mixture was refluxed at 75°C overnight under pure nitrogen protection. The reaction mixture was then cooled to room temperature and precipitated in 500mL of hexanes. The particles were then recovered by centrifugation at 3000rpm for 5 minutes and redispersed in THF using sonication and precipitated in hexanes again. This was repeated twice. The surface concentration of the silanes was found by thermogravimetric analysis to be ~1 molecule/nm^2.

Surface Energy Measurement: The static contact angles of water (γ_L =72 mJ/m^2, γ_L^d=21.8 mJ/m^2,γ_L^p=51 mJ/m^2), formamide (γ_L =58 mJ/m^2, γ_L^d=39 mJ/m^2,γ_L^p=19mJ/m^2) and diiodomethane (γ_L =50.8 mJ/m^2, γ_L^d=50.8 mJ/m^2,γ_L^p=0 mJ/m^2) on the functionalized surfaces were obtained by using the DropImage software on the Rame-Hart Goniometer. The contact angles were then used to evaluate the surface energies of these modified silicas using the Owens-Wendt method.[8] Surface energy is calculated by solving the following equation for two known liquids. It is based on Young Equation and the Good-Girifalco's geometric mean approximation of interaction energies.[9,10]

$$cos\theta = -1 + 2\frac{\sqrt{\gamma_S^d \gamma_L^d}}{\gamma_L} + 2\frac{\sqrt{\gamma_S^p \gamma_L^p}}{\gamma_L} \quad (1)$$

where γ_S^d and γ_L^d are the dispersive components of the solid and liquid surface energies respectively; γ_S^p and γ_L^p are the polar components of the solid and liquid respectively.γ_L is the total filler surface energy, given by the sum of γ_L^d and γ_L^p. The surface energies of the modified fillers (Amino, Octyl and Chloro silane modified silica) shown in Table 1, are average values calculated from the Owens-Wendt equation, solving for 2 liquids at a time. It was assumed that at our length scales the surface energy of a molecular cluster is similar to that of a planar macroscopic surface.[11] The surface energies of the matrix polymers in Table

1, were obtained by MQSPR prediction. These energies were used to calculate $\cos\theta$, ΔW_a, and W_s(shown in Table 2).

Table 1. Surface Energies of Surface Functionalized Silica and Polymer Matrix

Filler	**Surface Energy γ_f** **[mJ/m^2]**	**Dispersive γ_f^d** **[mJ/m^2]**	**Polar γ_f^p** **[mJ/m^2]**
Amino func. Silica	43.64	37.85	5.79
Octyl func. Silica	32.30	29.16	3.14
Chloro func. Silica	36.21	30.45	5.76
Polymer	**Surface Energy γ_p** **[mJ/m^2]**	**Dispersive γ_p^d** **[mJ/m^2]**	**Polar γ_p^p** **[mJ/m^2]**
Polystyrene	48.32	44.89	3.43
Poly(Methyl Methacrylate)	36.51	32.44	4.07
Poly(Ethyl Methacrylate)	33.02	31	2.02
Poly(2-Vinyl Pyridine)	48.56	42.93	5.63

Preparation of Nanocomposites: Nanoparticles in anhydrous THF were mixed with a 5% anhydrous THF solution of matrix polymer. This mixture was sonicated using a Sonics and Materials Vibracell VCX 600 Watt sonicator with a stepped microtip. Sonicated solutions were poured into clean aluminum boats and the solvent was driven off in a clean oven at 80^0C and vacuum dried at 120^0C for 16 hours. The samples were then pressed into dog-bone shaped samples and annealed at 140^0C (PS, PMMA, P2VP) and 110^0C (PEMA) for 24 hours.

*Microscopy and Image Analysis***:** Nanocomposite samples were embedded in an epoxy resin and microtomed into 60-80 nm thick slices using a diamond knife. These sections were transferred onto a copper grid and imaged using a JEOL-2010 transmission electron microscope (TEM).

The micrographs thus obtained were processed to remove noise and reduced to a resolution of 300x300 pixels. These images were binarized (black and white) using a 2 step process that ensures that the number of particles present is equal to the experimental volume fraction. After image binarization was completed, clusters were indicated on the pixilated 2D image.

The radius of each cluster (r_c) and the intercluster distance (r_d) between every cluster pairwas calculated and the averageof these values from two different images gaver_c and $r_d$in pixels (Note: Only one image is used in Amino-PEMA). It should be noted that r_c is taken as the average of radius of each filler weighed by the area of the filler in the 2D binary images, such that the existence of single dispersed fillers will not largely lower the r_c.These $r_c$and $r_d$values in pixels were then be converted to nanometers, using the scale bar. Calculated $r_c$and $r_d$values are shown in Table 2.

Glass Transition Temperature: The T_gs of nanocomposite samples were measured using Differential Scanning Calorimetry (DSC) and Dynamic Mechanical Thermal Analysis (DMTA). On the DSC, samples were subjected to three 20°C to 130°C cycles at a ramp rate of 10°C/min. The T_g values reported are averages from the second and third heating cycles. Temperature sweeps on the DMTA were run from 35°C to 150°C at a frequency of 1Hz. The T_g measured is the temperature at the peak loss factor (tanδ).

Table 2. Calculated values of energetic parameters, r_c, r_d and measured values of ΔT_g

System	Cos(θ)	W_{PF}/W_{FF}	ΔW_a [mJ/m^2]	W_s [mJ/m^2]	r_d-8 wt% [nm]	ΔT_g-8 wt% [ºC]	r_d -3 wt% [nm]	ΔT_g-3 wt% [ºC]
Chloro PMMA	1.00	1.05	0.62	-4.29	278.46	-4.58	190.85	-3.72
Amino P2VP	1.00	1.09	0.85	-7.90	156.13	-8.93	193.58	-6.49
Octyl PMMA	1.00	1.12	0.95	-8.51	291.86	-5.78	418.63	-5.50
Octyl PS	1.00	1.12	3.48	-11.57	266.71	-0.62	462.30	-1.58
Chloro PS	1.00	1.04	6.16	-8.86	374.69	-2.38	422.74	-1.33
Chloro PEMA	0.96	0.98	3.13	-0.45	449.31	-1.65	595.94	-0.88
Amino PS	0.92	0.95	4.71	-0.71	714.71	-0.47	974.90	-0.70
Amino PMMA	0.91	0.96	0.24	3.31	902.96	-0.84	1610.60	2.37
Amino PEMA	0.79	0.90	2.57	-0.08	1039	-0.35	1666.10	-0.095

Finite Element Analysis (FEA):

*Construction of 2-point correlation function***:** In order to reconstruct and/or generate the dispersion morphology in 3D, for the FEA of model composites, via predicted surface energy parameters, a 2-point correlation function was utilized.[12]The histogram of the 2-point correlation probability was measured in the binary images obtained through the image analysis process. The following expression was heuristically built to fit with the 2-point correlation probabilities, considering the necessary boundary conditions prescribed for a 2-point correlation function byJiao et al.[12]

$$f(x) = VF exp\left(\frac{-2x}{r_c}\right) + VF^2\left[\tanh\left(\frac{5\pi x}{2r_d}\right) + exp\left(\frac{-x}{r_d}\right)\sin\left(\frac{5\pi x}{2r_d}\right)\right] \quad (2)$$

*Heuristic-empirical expressions for r_c and r_d***:** Although the energetics of filler aggregation in glassy polymer nanocomposites, have been qualitatively explained in literature and in our paper, analytical quantitative expressions have rarely been formulated. We used a heuristic

method to construct empirical expressions to enable *a priori* prediction. The order of the *VF* term in the expression for r_c (i.e. the dependence of r_c on volume fraction) was determined by comparing the $r_c$at different loadings (3 and 8 weight %) for the same composite system. Using the same methodology and considering the work of Tong et al.,[13] we determined the form of *VF* in the expression of r_d. The dependence of $r_c$and r_d on surface energy was heuristically characterized through functions of a energy correlation variable x_{corr}. The fitting parameters in these functions were determined through an optimization process that minimized the difference between measured and predicted $r_c$and r_d values.

*FE simulation of DMTA temperature sweeps***:** FEA was performed using the ABAQUS software package. 3D representative volume element (RVE) geometries of PNCs with 30x30x30 hexahedron elements were reconstructed using the predicted 2-point correlation function. One XoY surface was fixed and the other was assigned with an oscillatory displacement boundary with frequency of 1Hz. This is consistent with the loading frequency used in the DMTA experiment. The properties of the polymer matrix were obtained from the experimentally determined Prony series at reference temperature and the corresponding shifting factors required for time-temperature superposition. External python scripts for ABAQUS were developed to update the Prony series of polymer matrix and interphase at different temperatures and to control the simulation procedure. The complex tensile modulus was obtained from the output of the simulation at each temperature. The final output curves for $E'/E''/\tan\delta$ over T, were then used to predict ΔT_g.

Web Tool:

After being validated using best-practices methods,[2] both the dispersive and polar bootstrap aggregate polymer surface energy models were incorporated into a web-based toolkit to facilitate their use by the broader community, and to provide a front end for our hybrid MQSPR/Finite Element Analysis (FEA) modeling tool. Currently, the MQSPR portion of the

tool allows users to compute surface energy components for either existing or designed polymers, which can then be used together with the predictedsurface energy components of the selected nanofiller as part of a larger web-based hybrid tool discussed at the end of the paper.In practice, the web-based toolkit is divided into three major components across two sites:

The first provides the user with a simplified front end for designating polymer matrix structures, nanoparticulate core type (silica is the only option at this point) and volume fraction, along with a drop-down menu of nanoparticle surface functionalizations. While stepping through the web-based tool, the user receives information on surface energies, dispersion paramaters and eventually thermomechanical properties. See: http://reccr.chem.rpi.edu/polymerizer to utilize the webtool.

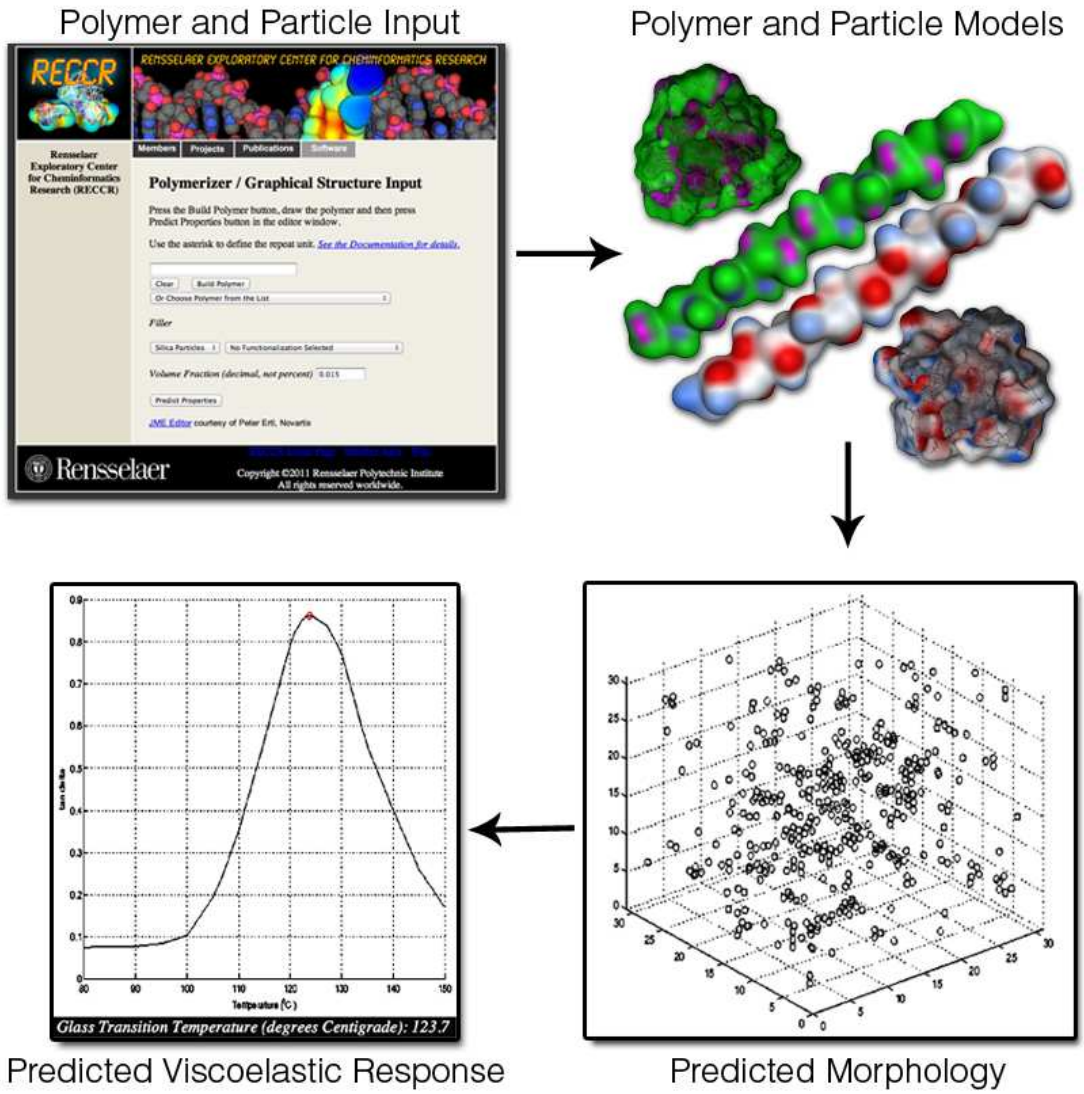

Figure 1. Actual screenshots of the Polymerizer MQSPR/FEA nanocomposite prediction webtool input and several stages of its output. The workflow of the multi-scale tool is analogous to the flowchart shown Figure 1, which illustrates the process of defining the polymer matrix, particle functionality and volume fraction, leading to predictions of surface energies, the resulting 3D particle distributions, and the tanδ curve.

Image Processing:

Figure1: was prepared using Adobe Photoshop.

Figure 2: a.Images are screenshots from MOE and b.Plots are results plotted by the MQSPR tool.

Figure 3: The graph was plotted using GNUplot and the labels were added on Adobe Photoshop

Figure 4: The image was prepared using Adobe Photoshop. The inset micrographs were obtained using Transmission Electron Microscopy.

Figure 5.a, b and c: were plotted on OriginPro and d. is a screenshot from the FEA environment.

Figure 6: a and b were plotted using Origin Pro.

Supporting References

[1]MOE (The Molecular Operating Environment) Version 2009.10, software available from Chemical Computing Group Inc., 1010 Sherbrooke Street West, Suite 910, Montreal, Canada H3A 2R7. http://www.chemcomp.com

[2]C. M. Breneman, T. R. Thompson, M. Rhem, M. Dung, *Computers & Chemistry***1995**, *19*, 161–179.

[3]Labute, P. An Integrated Application in MOE for the Visualization and Analysis of Protein Active Sites with Molecular Surfaces, Contact Statistics and Electrostatic Maps. *J. Chem. Comput. Group***2006**.

[4]Santavy, M.; Labute, P. Electrostatic Fields and Surfaces in MOE *J. Chem. Comput. Group* [Online], **1998**. http://www.chemcomp.com/journal/grid.htm (accessed 9/20/2011).

[5]R Development Core Team, R: *A Language and Environment for Statistical Computing*. Vienna, Austria, **2005**.; Mevik, B.; Wehrens, R. The Pls Package: Principal Component and Partial Least Squares Regression in R. J. Stat. Soft. 2007, 18 (2), 1-24.

[6]Chih-Chung Chang and Chih-Jen Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, **2011**. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7]Meyer, D. Support Vector Machines: The Interface to Libsvm in Package E1071. *R News***2001**, *1*, 23-26.

[8]D. K. Owens, R. C. Wendt, *Journal of Applied Polymer Science***1969**, *13*, 1741–1747.

[9]R. Good, L. Girifalco, *The Journal of Physical Chemistry***1960**, *64*, 561–565.

[10]F. M. Fowkes, *Industrial & Engineering Chemistry***1964**, *56*, 40.

[11]O. Sinanoğlu, *The Journal of Chemical Physics***1981**, *75*, 463.

[12]Y. Jiao, F. Stillinger, S. Torquato, *Physical Review E***2007**, *76*, 1–15.

[13]W. S. Tong, J. M. Rickman, K. Barmak, *Acta Materialia***1999**, *47*, 435–445.