Electronic Supplementary Material (ESI) for Digital Discovery.
This journal is © The Royal Society of Chemistry 2024

Data-driven representative models to accelerate atomistic simulations of bitumen and biobased complex fluids.

Daniel York,[a] Isaac Vidal-Daza,[a,b] Cristina Segura,[c] Jose Norambuena-Contreras,[d] Francisco J. Martin-Martinez [a*]

[a] *Department of Chemistry, Swansea University, Swansea, SA2 8PP, UK*

[b] *Grupo de modelización y diseño molecular, Universidad de Granada, Granada, 18071, Spain*

[c] *Unidad de Desarrollo Tecnológico, Universidad de Concepción, Coronel 4191996, Chile*

[d] *LabMAT, Department of Civil and Environmental Engineering, University of Bío-Bío, Concepción 4051381, Chile*

Supplementary material

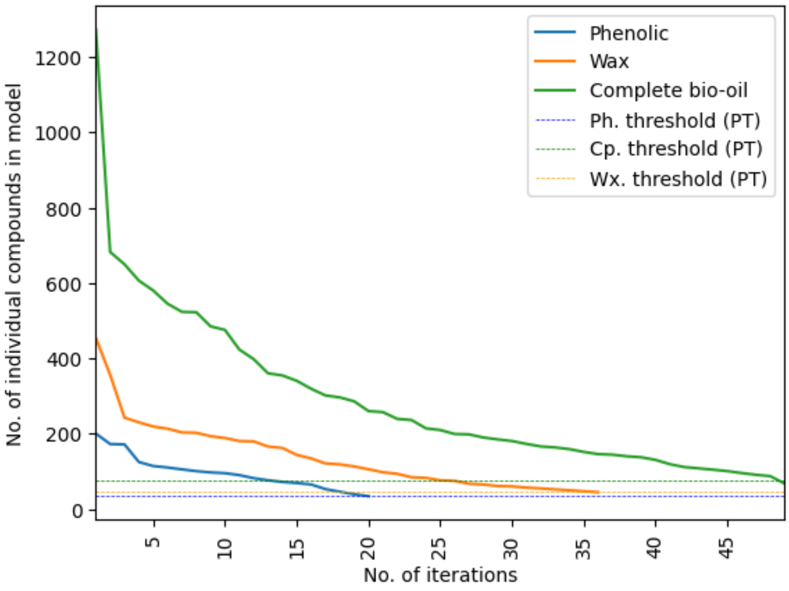
1. Proportional representation algorithm (PT model) – see Abundancy-based model generation system in main text

The algorithm for generating the PT models is given by **equation (s1)** (*equation (9) in the main text)*, where a_i is the abundance of the compound in question, a_s is the sum of all abundancies and N is the total number of characterized compounds.

$$\frac{a_i/a_s}{\sum_{i=1}^{N} a_i/a_s} \geq \mathbf{1/N} \tag{s1}$$

Equation (s1) results in a dynamic selection rule that is specific to a given in mixture – in our case, Py-GCMS data from the pyrolysis of pine-bark- where, the abundancy of each compound is divided by the smallest abundancy in the Py-GCMS data (normalised). This yields a stoichiometric count of each unique compound in the bio-oil and, to allow the selection rule to dynamic between mixtures, the number of selected compounds is equal to the number of unique compounds in the mixture (i.e 75, 36 and 48 for the complete bio-oil and phenolic/wax fractions respectively).

The selection process the PT method implements is iterative; where the sum of normalised abundancies is calculated and if the sum is less than or equal to the number of unique compounds in the mixture (i.e. 75 in the complete bio-oil), that group of compounds. However, if the sum is greater than the number of unique compounds in the mixture, the least abundant compound is removed from consideration and the process started all over again. This results in an elimination process akin to instant runoff voting, where the lowest ranked candidate is eliminated, and votes recalculated. In the case of pine bark, the number of iterations is equal to the number of omitted compounds with a larger sample of molecules requiring more iterations **fig. s1,** in this case the phenolic, wax, and complete models took 20, 36 and 49 iterations respectively. This resulted in the definition of a unique selection threshold for each fraction: 1.09% (complete model), 1.57% (wax fraction) and 2.08%



(phenolic fraction) with these molecular models being comprised of 17/36 (47%), 12/47 (26%) and 27/75 (36%) compounds for the respective PT models of the phenolic fraction, wax fraction, and complete bio-oil respectively.

***Fig. s1** The number of iterations required by the **PT** methodology to select compounds for molecular models.*

This method can be considered a type of stratified sampling where an abstract box is created, and a single molecule of the least abundant compound is placed. This is followed by placing an amount of each other compound in this abstract box proportional to the least abundant compound. For example, the complete bio-oil contains 75 unique compounds and if 75 individual compounds were picked out of the box, the most likely combination of selected compounds is given by the **PT** selection algorithm - where multiple instances of each unique compound can be selected for the model – i.e. leaving the least abundant compounds in the box.

2. Disproportionate influence

The proportion of each compound in the final models (P_i) was defined using **equation (s2)** (*equations (4) in the main text)*, where a_i is the abundancy of a compound in the model and $\sum a_{selected}$ is the summative abundance of the selected compounds.

$$P_i = a_i/\sum a_{selected} \tag{s2}$$

Equation (s2) was used to calculate values of P_i of all 4 models (FT, PT, AG and SG) and **fig. s2** the performance of each of these models versus the all-molecule benchmarks for the complete bio-oil and the phenolic/wax fractions.

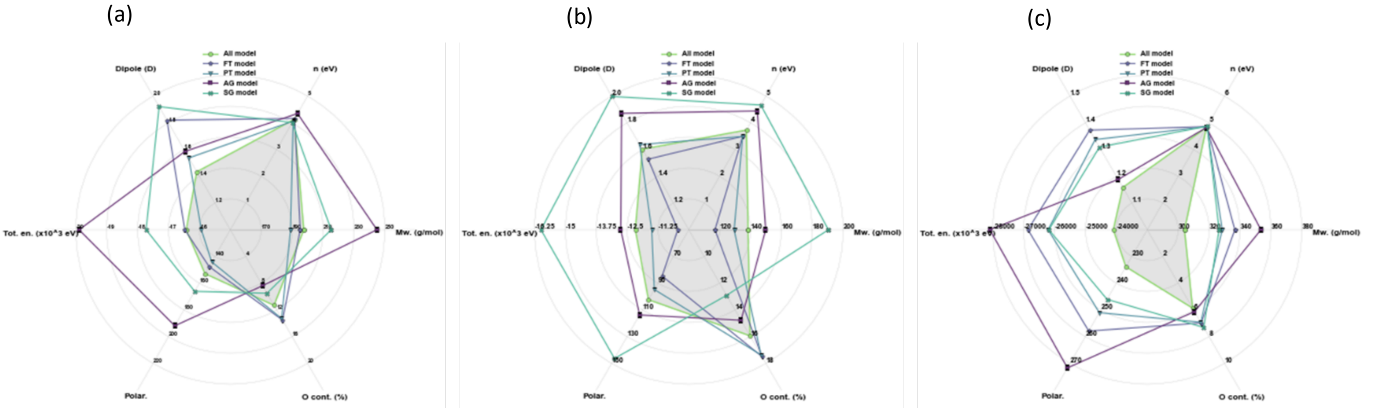


***Fig. s2** Average descriptors for each model (a) complete bio-oil, (b) phenolic fraction and (c) wax fraction. The average descriptors of each method to generate molecular models are shown in (a-c), the FT (indigo line), the PT method (light blue line), the AG method (deep purple line), and the SG method (jade green line) together with the all-molecule model (light green line).*

The molecular-classification methods perform the worst versus the all-molecule model in **fig. s2** and this is because the AG and SG models only select a single compound from a molecular class. This can cause the proportion of that molecular class within a mixture to be misrepresented when P_i is calculated as above with **equation (s2)**. This is because the proportion of the selected compound within a mixture is independent of the overall molecular class and as an example, the molecular class that consists of multi-ring structures for 1.87%, 1.11% and 3.40% in the complete bio-oil, phenolic and wax fraction, but the compound selected from this class in the AG model accounts for 4.88%, 6.56% and 8.76% of final model composition. In this case, the proportion of this molecular class is overestimated in each case and is exasperated as the multi-ring structure group doesn't meet any selection threshold in the FT and PT methods. This causes the stoichiometry of compounds selected in the molecular-classification models to be inaccurate versus experimental data when the stoichiometry is derived from **equation (s2)** and consequently, the molecular-classification models perform inconsistently. Another example of this disproportionate influence can be seen in the compound selected from the phenolic molecular class for the AG/SG models of the phenolic fraction as the selected compound has an abundancy (in Py-GCMS data) of 9.89% in the AG model, but only 1.86% in the SG model. When these abundancy values are so different for a compound from the same molecular class between methods, the calculated average descriptors are vastly different between models as molecular classes vary differently in their overall proportions between models.

As described in the main text, for the molecular-classification system, a different approach to defining the proportions of compounds in the AG or SG models was required based on this lacklustre performance seen in **fig. s2**. This approach is shown in **equation (s3)** (*equation (5) in the main text*) where $\sum a_{subclass}$ is the summative abundancy of all the compounds within that compounds subclass, $\sum a_{all\ compounds}$ is the summative abundancy of all compounds in the Py-GCMS data and P_i is the calculated proportion of the compound in question. Essentially, each compound is a representative of its molecular class and its proportion equal to proportion of that molecular class, not the compounds individual proportion.

$$P_i = \sum a_{subclass} / \sum a_{all\ compounds} \tag{s3}$$

The idea in defining proportions this way mitigates any disproportionate influence when calculating the weighted averages of DFT descriptors. The results of using both **equations (s2)** and **(s3)** for the AG model are shown in **table s1** and the results of using these equations for the SG model are shown in **table s2**. These tables include the benchmark values as calculated from the all-molecules model and then the subsequent values of average descriptors calculated using **equation (s2)** – where there is disproportionate influence - and **equation (s3)** – where any disproportionate influence has been mitigated and weighted averages more closely match the benchmark values of the all-molecule model.

Table s1 *Average descriptors from each complete model for the entire bio-oil and both the phenolic and wax fractions. Ph./Wx. and Cp. denote the phenolic/wax fractions, and complete bio-oil. Fixed corresponds to the average descriptors calculated after any disproportionate influence has been mitigated using equation (s3) and Disp. Inf. Corresponds to average descriptors before this effect was accounted for and these were calculated using equation (s2).*

	Model	Mw (g/mol)	η (eV)	Polarizability	Dipole moment	Total energy (eV)	Oxygen content (%)
Cp.	**Benchmark**	197.55	4.13	152.81	1.43	-16426.84	11.30
	Fixed (AG)	210.43	4.23	165.15	1.51	-17157.38	8.36
	Disp. Inf. (AG)	244.89	4.35	191.86	1.59	-17157.38	8.37
Ph.	**Benchmark**	138.38	3.74	102.27	1.60	-12165.78	17.93
	Fixed (AG)	123.88	3.85	93.16	1.77	-10681.57	13.39
	Disp. Inf. (AG)	149.87	3.89	113.74	1.87	-127542.36	13.39
Wx.	**Benchmark**	304.07	4.82	243.85	1.15	-24098.27	5.86
	Fixed (AG)	347.94	4.81	280.68	1.06	-27439.l75	6.12
	Disp. Inf. (AG)	353.57	4.82	281.68	1.19	-28093.74	6.12

Table s2 *Average descriptors from each complete model for the entire bio-oil and both the phenolic and wax fractions. Ph./Wx. and Cp. denote the phenolic/wax fractions, and complete bio-oil. Fixed corresponds to the average descriptors calculated after any disproportionate influence has been mitigated and Disp. Inf. Corresponds to average descriptors before this was calculated.*

	Model	Mw (g/mol)	η (eV)	Polarizability	Dipole moment	Total energy (eV)	Oxygen content (%)
Cp.	**Benchmark**	197.55	4.13	152.81	1.43	-16426.84	11.30
	Fixed (SG)	193.40	4.15	151.78	1.54	-15860.83	9.50
	Disp. Inf. (SG)	215.27	4.35	166.10	1.92	-17723.87	9.50
Ph.	**Benchmark**	138.38	3.74	102.27	1.60	-12165.78	17.93
	Fixed (SG)	136.06	3.76	104.47	1.78	-11608.90	12.45
	Disp. Inf. (SG)	190.51	3.93	146.31	2.0	-15989.32	12.45
Wx.	**Benchmark**	304.07	4.82	243.85	1.15	-24098.27	5.86
	Fixed (SG)	296.58	4.82	237.53	1.07	-23562.46	7.31
	Disp. Inf. (SG)	326.49	4.87	256.03	1.31	-26184.28	7.31

3. Other distributions of descriptors in the models

The distribution of reactivity for all 3 mixtures, complete bio-oil and the phenolic/wax fractions was included in the main text and is the most important for our uses in describing the reactivity of a mixture. However, the distributions of other DFT descriptors were also generated for each mixture – excluding the total energy as this distribution is proportional to the distribution of the molecular weight. **Figs. s3-s5** show histograms with the distribution of values for molecular weight in the all-molecule model and the phenolic/wax fractions.

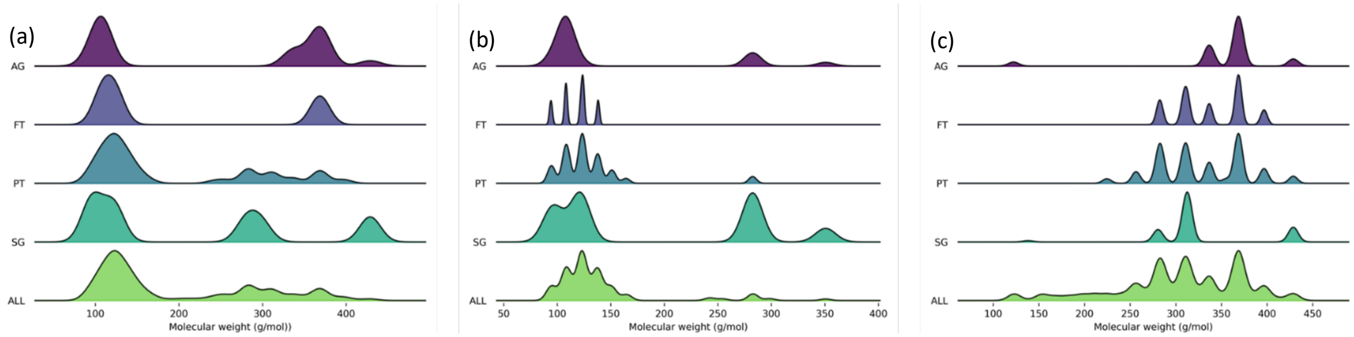


Fig. s3 *Distribution of molecular weight in molecular models for (a) complete bio-oil, (b) the phenolic fraction and (c) the wax fraction of pine bark-derived bio-oil.*

Fig. s4 shows histograms with the distribution of values for dipole moment in the all-molecule model and the phenolic/wax fractions.

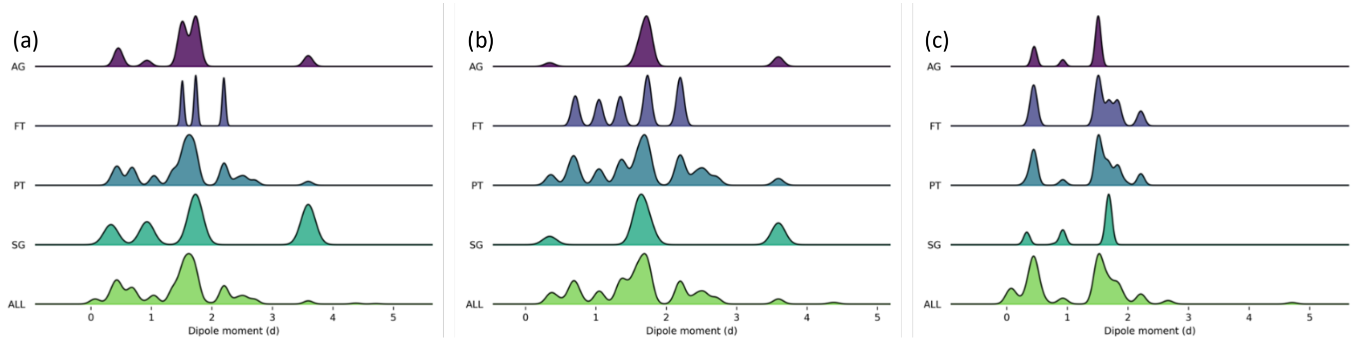


Fig. s4 *Distribution of molecular weight in dipole moment models for (a) complete bio-oil, (b) the phenolic fraction and (c) the wax fraction of pine bark-derived bio-oil.*

Fig. s5 shows histograms with the distribution of values for polarizability in the all-molecule model and the phenolic/wax fractions.

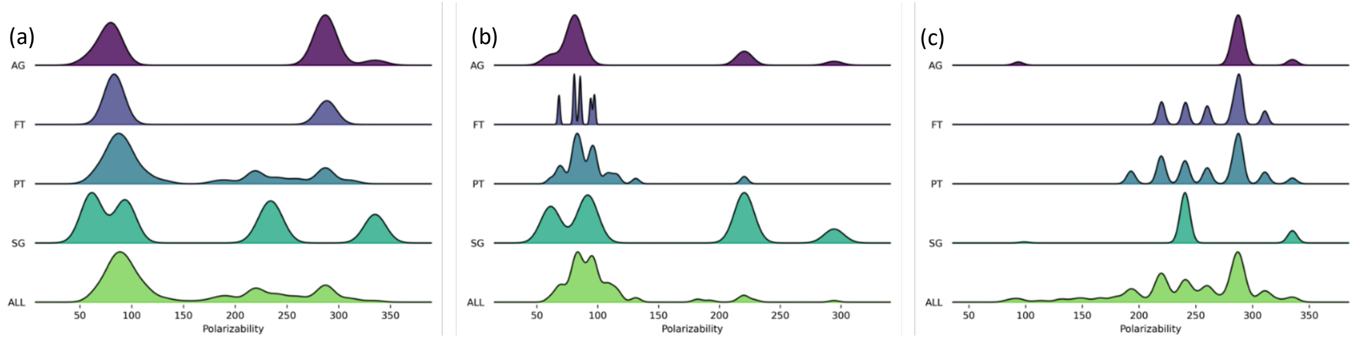


Fig. s5 *Distribution of polarizability in dipole moment models for (a) complete bio-oil, (b) the phenolic fraction and (c) the wax fraction of pine bark-derived bio-oil.*