

Supporting Information

On the design of optimal computer experiments to model solvent effects on reaction kinetics

Lingfeng Gui^a, Alan Armstrong^b, Amparo Galindo^a, Fareed Bhasha Sayyed^c, Stanley P. Kolis^d, Claire S. Adjiman^{a,*}

^a*Department of Chemical Engineering, The Sargent Centre for Process Systems Engineering and Institute for Molecular Science and Engineering, Imperial College London, London, SW7 2AZ, UK*

^b*Department of Chemistry and Institute for Molecular Science and Engineering, Imperial College London, Molecular Sciences Research Hub, White City Campus, London, W12 0BZ, UK*

^c*Synthetic Molecule Design and Development, Eli Lilly Services India Pvt Ltd, Devarabeesanahalli, Bengaluru, 560103, India*

^d*Synthetic Molecule Design and Development, Eli Lilly and Company, Lilly Corporate Center, Indianapolis, 46285, Indiana, United States*

***Corresponding author:** Claire S. Adjiman, *Email: c.adjiman@imperial.ac.uk.

S1. LINEAR FREE ENERGY RELATIONSHIP (LFER)

The LFER, trained using a 9-solvent set selected based on the D-optimality criterion for each selection space, is shown below:

Selection space 1:

$$\ln k^{\text{L,LFER}} = -16.82 + 6.75A + 4.68B + 0.18n^2 + 6.34\gamma + 0.95\epsilon - 0.19\phi + 0.29\psi$$

Selection space 2:

$$\ln k^{\text{L,LFER}} = 0.27 + 10.42A + 5.32B - 7.17n^2 + 7.68\gamma - 1.55\epsilon - 1.50\phi + 1.60\psi$$

Selection space 3:

$$\ln k^{\text{L,LFER}} = -20.04 + 0.29A + 1.59B + 2.77n^2 + 2.95\gamma + 14.68\epsilon - 1.52\phi + 1.13\psi$$

Selection space 4:

$$\ln k^{\text{L,LFER}} = -20.11 + 1.29A + 0.45B - 0.16n^2 + 2.55\gamma + 6.03\epsilon - 0.30\phi - 0.50\psi$$

The impact of specific solvent descriptors/properties on reaction kinetics can be inferred from the coefficients associated with these LFERs, although, as in most data-driven models, it can be difficult to assign physical meaning to the coefficients. The regression coefficients in these LFER models indicated that solvents with high hydrogen bond acidity, hydrogen bond basicity, and surface tension consistently lead to increased rate constants, and thus, higher solvent rankings. Such a favourable effect is to be expected for solvents that can form hydrogen bonds (as donor or acceptor), as such solvents can significantly stabilise the transition state of the Menschutkin reaction, where charge separation occurs, relative to the neutral reactants.

S2. LOCATIONS OF SUPPLEMENTARY MATERIALS IN THE ZENODO REPOSITORY AND THE INFORMATION CONTAINED

Address of the Zenodo online repository: [10.5281/zenodo.8396100](https://zenodo.org/doi/10.5281/zenodo.8396100)

- `supplementary_material_v3` → `chapter6` → `Fedorov` → `Fedorov_SS1.ipynb`:
This Jupyter notebook implements Fedorov's algorithm for selection space 1.
- `supplementary_material_v3` → `chapter6` → `Fedorov` → `Fedorov_SS2.ipynb`:
This Jupyter notebook implements Fedorov's algorithm for selection space 2.
- `supplementary_material_v3` → `chapter6` → `Fedorov` → `Fedorov_SS3.ipynb`:
This Jupyter notebook implements Fedorov's algorithm for selection space 3.
- `supplementary_material_v3` → `chapter6` → `Fedorov` → `SS1`: This text file lists the names of the solvents in selection space 1 along with their solvent descriptor values.
- `supplementary_material_v3` → `chapter6` → `Fedorov` → `SS2`: This text file lists the names of the solvents in selection space 2 along with their solvent descriptor values.
- `supplementary_material_v3` → `chapter6` → `Fedorov` → `SS3`: This text file lists the names of the solvents in selection space 3 along with their solvent descriptor values.
- `supplementary_material_v3` → `chapter6` → `MBDoE_GAMS` → `GAMS_SS1.gms`: This GAMS code contains the mixed-integer non-linear programming formulation of the MBDoE problem for solvent selection from selection space 1, based on the D-optimality criterion.
- `supplementary_material_v3` → `chapter6` → `MBDoE_GAMS` → `GAMS_SS1.lst`: This is the output file generated by running `GAMS_SS1.gms`. The optimal solvent set is located starting from line 571506.
- `supplementary_material_v3` → `chapter6` → `MBDoE_GAMS` → `GAMS_SS4.gms`: This GAMS code contains the non-linear programming formulation of the MBDoE problem for solvent selection from selection space 4, based on the D-optimality criterion. Lines 72 – 85 detail the bounds on solvent descriptors, same as those used for generating selection space 3.

- supplementary_material_v3 → chapter6 → MBDoE_GAMS → GAMS_SS4.lst: This is the output file generated by running GAMS_SS4.gms, the optimal solvent set is located starting from line 915.
- supplementary_material_v3 → chapter6 → MBDoE_GAMS → initial.inc: This text file lists the initial guess of the optimal solvent set for the MBDoE problem of selection space 1. The file is read by GAMS_SS1.gms.
- supplementary_material_v3 → chapter6 → regression.xlsx: This Excel file provides the data for Figure 9 in the manuscript. The file is organised as follows:
 - Sheet 1: Contains linear regression results from the optimal solvent set of various sizes using the Minnesota solvent descriptor database for training and the CAMD design space for validation.
 - Sheet 2: Contains quadratic regression results from the optimal solvent set of various sizes using the Minnesota solvent descriptor database for training and the CAMD design space for validation.
 - Sheet 3: Contains linear regression results from the optimal solvent set of various sizes using the CAMD design space for both training and validation.
 - Sheet 4: Contains quadratic regression results from the optimal solvent set of various sizes using the CAMD design space for both training and validation.
 - Sheet 5: Contains the linear regression results corresponding to the parity plots in Figure 7 of the manuscript, including the intercept and coefficients of each LFER in section 3.4.
- supplementary_material_v3 → chapter6 → solvents.xlsx: This Excel file lists the solvents within each selection space. The sheet named as “MBDoE solvents” contains the optimal 9-solvent sets for all selection spaces, the optimal 13-solvent sets for SS1 and SS2, and the optimal 49 solvent sets for SS1 and SS2, corresponding to the data presented in Figure 8 of the manuscript.
- supplementary_material_v3 → chapter6 → SS2_generator.gms: This GAMS code contains the mixed-integer linear programming (MILP) formulation used

to generate selection space 2. Atom groups used for generating SS2 are defined between line 11 and 20. Lines 360 to 404 detail the constraints on the maximum number of each atom group allowed in the designed molecule, with some atom groups deactivated (i.e., the maximum number set to 0). The “equations” section of the GAMS code outlines all constraints used in generating of SS2. For an explanation of each constraint, please refer to the paper at: <https://doi.org/10.1016/j.compchemeng.2023.108345>.

S3. RELATIONSHIP BETWEEN MODEL PERFORMANCE AND D-OPTIMALITY CRITERION VALUES FOR THE CYCLISATION OF OXYMA/DIC ADDUCT

We have also applied the MBD_{oE} approach to the cyclisation of the adduct of ethyl (hydroxyimino)cynoacetate (Oxy_{ma}) and diisopropylcarbodiimide (DIC), which produces hydrogen cyanide (HCN) in peptide synthesis (referred to as “the HCN reaction”, the reaction is described in the papers <https://doi.org/10.1016/B978-0-323-95879-0.50102-8> and <https://doi.org/10.1016/j.compchemeng.2023.108345>). The test set for the HCN reaction is eight solvents commonly found in a chemical lab, i.e., toluene, chlorobenzene, ethyl acetate, tetrahydrofuran, acetone, acetonitrile and nitromethane. The resulting probability distributions of mean absolute deviations (MADs) and rank correlations (RCs) over the D-optimality criterion values for the HCN reaction are shown in Figure S1 and S2, respectively. It shows that for the HCN reaction, greater D-optimality criterion values generally lead to larger probability of obtaining linear free energy relationships with MADs smaller than 3 log units. When the natural logarithm of the D-optimality criterion value is greater than -2, the probability of achieving a MAD smaller than 3 log units is above 80%. As for RC, a trend can also be observed that greater D-optimality criterion values result in larger probability of obtaining a model with RC greater than 0.7. However, the maximum probability that can be achieved is only between 60% and 70%. These results are, in general, consistent with those obtained for the Menschutkin reaction in the manuscript.

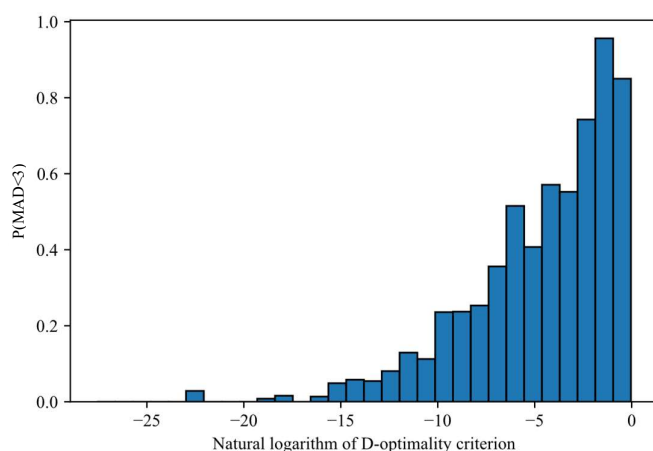


Fig. S1 Probability distribution of obtaining linear free energy relationships with a MAD < 3 log units for the HCN reaction.

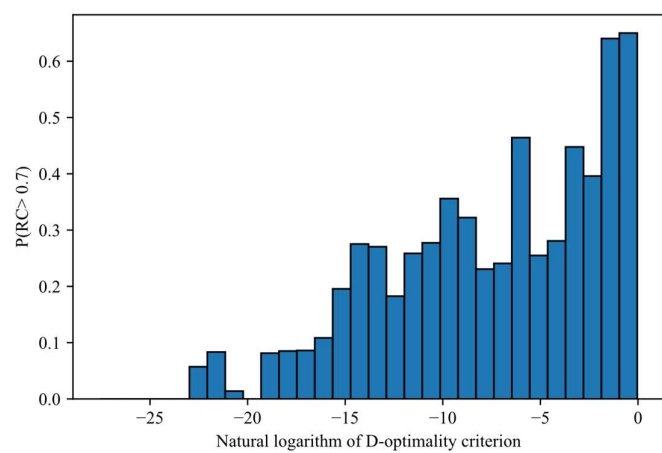


Fig. S2 Probability distribution of obtaining linear free energy relationships with a $RC > 0.7$ for the HCN reaction.