BSc. Industrial Biochemistry
University of Limerick
2021

Modelling the Crystallization Kinetics of Organic Compounds Using Artificial Neural Networking

Date Submitted: 01/04/2021

Author: Anthony Maguire
Student ID: 17237475
Supervisor: Vasanth Kannuchamy

Modelling the Crystallization Kinetics of Organic Compounds Using Artificial Neural Networking†

Anthony Maguire†

A feed-forward back propagation artificial neural network (ANN) was constructed and tested to predict the linear, G, and overall, R_g, crystal growth kinetics from different polymorphic forms of Acetaminophen, Carbamazepine, Curcumin, Fenofibrate, Piracetam, Risperidone, Salicylamide, Salicylic Acid and Tolbutamide, under different operating conditions. The operating conditions for each of the organic compounds were used as inputs to predict G and R_g values. The ANN model was established to be accurate in modelling the crystal growth kinetics for multiple organic compounds simultaneously. Furthermore, the constructed ANN model was determined to be precise in predicting G and R_g values for a new data set, separate from the trained ANN, which displayed the ANN model's efficacy for the determination of crystal growth kinetics of a compound, or multiple compounds, under varying operating conditions. The neural network's predicted linear, G_{ANN}, and overall, $R_{g,ANN}$, crystal growth rates were compared to those from a multiple non-linear regression (MNLR) analysis, G_{MNLR}, and $R_{g,MNLR}$. The ANN was observed to be more applicable for modelling G and R_g values with high accuracies in predicting the crystal growth kinetics. The correlation coefficient (r^2) between experimentally determined linear, G_{Exp}, and overall, $R_{g,Exp}$, crystal growth rates and the crystal growth rates predicted by the ANN, G_{ANN} and $R_{g,ANN}$, were measured to be 0.9970 and 0.9958, respectively. Alternatively, for the predicted crystal growth kinetics from the MNLR analysis, G_{MNLR} and $R_{g,MNLR}$, the r^2 values were determined to be 0.1248 and 0.1144, which demonstrated a poorly representative model with significant inaccuracies in predicting crystal growth kinetics. The r^2 values for the new data set between the experimentally determined crystal growth kinetics, G_{Exp} and $R_{g,Exp}$, and the ANN-predicted crystal growth kinetics, G_{ANN} and $R_{g,ANN}$, was observed to be 0.9959 and 0.9945.

1 Introduction

Crystallization has sustained itself as the primary separation and purification process in the pharmaceutical industry leading to further extensive research in crystallization operations to solve several critical factors that require increased control of the process involved. There is a weight of importance in understanding the fundamentals of crystal growth kinetics, the operating characteristics of crystallization equipment, and critical issues dealing with scale-up, to increase the control of a crystallization process. Understanding crystal properties such as solubility, supersaturation, morphology, and polymorphism is needed in order to optimize crystallization processes. Solubility data is required for determining the supersaturation and crystallization kinetics of a compound, and in some instances, is used to optimize overall synthesis of pharmaceuticals. Generation of supersaturation within a process is a requirement for crystallization to occur, which can be created by varying conditions such as temperature, agitation speed, solvent composition, and its salt form. In industrial applications, controlling the crystalline size is dependent on the control of nucleation and crystal growth. The complexity of the kinetics of crystal growth deems increasingly intricate due to factors relating to predicting polymorphism and isolating individual polymorphs of a compound.[1]

The complex process of crystallization involves two successive steps: nucleation and crystal growth. Nucleation is the first formation of the solid phase that occurs due to clustering and aggregation of molecules of ions,[2] in which assembly of these molecules are held together by intermolecular forces, H-bonds, π-π interactions, and van der Waals forces. This stage in crystallization is traditionally described through the classical nucleation theory.[3] Crystal growth is a diffusion and integration process involving the transport from a solution to the crystal surface and the surface integration of the crystal lattice, resulting in the growth of crystals. Modelling of the crystal growth kinetics process is normally represented by a two-step crystal growth model, or a semi-empirical expression denoting the overall growth kinetics. Simulation of the crystal growth process using a semi-empirical expression, as mentioned above, can be useful but limitations arise in the inability to generalize an expression correlating all operating variables involved in the process. This is due to variance in the order of reaction kinetics, in terms of the overall growth kinetics expression. There are several theoretical surface integration models proposed in order to evaluate the overall crystal growth kinetics. The Burton Cabrera Frank (BCF), and Birth and Spread (B+S) surface integration models[4] are adequate examples of theoretical models that are based on specific assumptions for the occurrence of crystal growth. The basis of the BCF Model assumes that growth occurs along such screw dislocations forming spiral staircases, whereas the B+S Model assumes new growth steps are initiated by 2D nucleation on the surface. Limitations of such semi-empirical expressions and accountability of all operating variables from theoretical considerations has created rationality to use artificial neural networks (ANN) as a preferable option for solving complex problems.

The ability for artificial neural networks to model non-linear and complex relationships between a data set, in addition to generalization of the model to predict unseen data, allows for this prediction technique to be viable in studying crystal growth kinetics. There are a number of fields that ANN has been applied to, including character recognition[5], image processing and visu-

alization[6], stock performance, and processes in chemical engineering. ANN are dissimilar to many other prediction techniques where it does not impose any restrictions on input variables, leading to improved models. The effectiveness of using ANN is visible in the contribution of excessive investment returns from a highly volatile dataset in Kimoto's prediction system for stocks on the Tokyo Stock Exchange.[7] In chemical engineering, the applicability of ANNs have been successful in predicting polymer quality in an industrial reactor's unit[8], modelling predictive control using a process model[9], the solubility of proteins[10], and modelling solid-liquid adsorption processes by pseudo second order kinetics.[11] The prediction of crystallization kinetics using ANNs has only been applied in a few studies such as Kumar, Martins and Rocha's analysis of the crystal growth kinetics of sucrose[12], in addition to Yang and Wei's application in predicting ciprofloxacin's agglomeration coefficient and crystallization kinetics.[13] C. Damour et al.[14] illustrated the advantages of applying ANN on developing an optimized process control strategy for industrial crystallization concluding the increased efficiency of this proposed strategy to improve process control.

Previously, crystallization studies have been heavily focused on inorganic compounds, although in recent years, there has been a slow transition towards focusing on the nucleation and crystal growth of organic molecules. Poor fundamental understanding in polymorphic systems has created increased complexity in terms of crystallization, but following technological advancements and discoveries, improvements in experimental techniques and analysis will allow for incremental increase in crystal growth efficiency and control of nucleation in the crystallization process. In this study, the crystal growth kinetics of nine different organic compounds (Figure 1), Acetaminophen (AAP), Carbamazepine (CBMZ), Curcumin (CUR), Fenofibrate (FF), Piracetam (PCM), Risperidone (RIS), Salicylamide (SM), Salicylic acid (SA) and Tolbutamide (TBM), in a number of solvents (discussed in Section 2 Database) is analysed and predicted by constructing an ANN model. The growth of crystals from organic molecules can be quite complex with different operating variables involved in the crystallization process such as, mean particle size, agitation speed, temperature, relative supersaturation, initial supersaturation, molecular volume, molecular weight, crystallographic density, crystal shape factors, solvent composition, lattice energies, and activation energies.

The ability of ANNs to correlate hidden relationships between inputs and outputs of a process deems it a suitable technique for predicting both the linear crystal growth rate, G (m/s) and overall crystal growth rate, R_g $(kg/m^2 s)$, of the compounds in Figure 1. The consistency of an ANN involves input and output layers that are connected by several nodes. In this case, construction of a multiple layer feed-forward back-propagation neural network is created to predict the crystallization kinetics of these organic compounds. The feed-forward aspect of an ANN can be generally explained through matrix multiplication, where input nodes are connected to specific weights. These weights transform the input data with the neural networks hidden layers. A bias is then added to the weight sum of the inputs and is then fed into an activation function, proceeding with interpretation of this data to produce outputs throughout the hidden and output layers. The process of back-propagation includes training the neural network throughout the individual layers by adjusting the weights to reduce the outputs error. There are numerous back-propagation algorithms used for training an ANN such as BFGS Algorithm $(trainbgf)$, Levenberg-Marquardt $(trainlm)$, and Automated / Bayesian Regularization $(trainbr)$.[15] Specifically, The Bayesian Regularization algorithm was used to train the ANN due to its profound ability to prevent overfitting and improve generalization of the network. The constructed ANN adjusts the input and output's transfer functions until convergence is reached. The applicability of the trained ANN to predict both linear and overall crystal growth rate of the nine organic compounds, was tested by using a new data set of inputs separate from the previous data set, ensuring no bias in prediction of the crystallization kinetics.

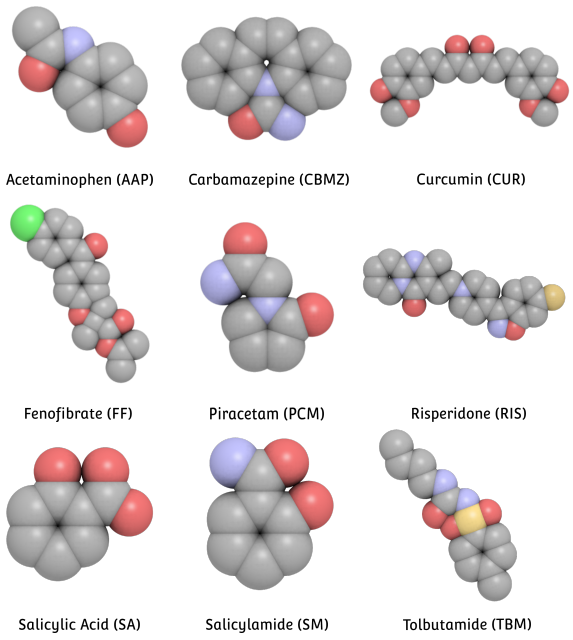


Fig. 1 Molecular visualizations of the nine organic compounds used for predicting crystallization kinetics using ANN (QuteMol).

2 Database

The key principle of creating an accurate ANN model is to employ the necessary operating variables needed as inputs, while maintaining uniformity of the units throughout the input and output layers for each organic compound in the neural network. Prior to procuring a database needed to train and test the ANN, it is required to understand the inputs utilized in creating the database. To illustrate the differences of the physical and chemical properties of the nine organic compounds, the following inputs were used to calculate the crystal growth kinetics: mean particle size (L_0), temperature (T), agitation speed (v), relative supersaturation $(S-1)$, initial supersaturation $(S-1)_i$, molecular volume (mV), molecular weight (mW), crystallographic density (ρ_c), volumetric shape factors (f_v), surface shape factors (f_s), solvent molecular weight (mW_s), lattice energy (E_l) and activation energy

(E_a). The data set used in this study was collected from a number of publications to ensure validity of the ANN model by providing sufficient data of a wide range of organic compounds with high variance in properties such as, crystal habit, shape factors, crystallographic density, molecular volume, molecular weight, lattice energy and solvent composition. The data needed to calculate the crystallization kinetics of these organic compounds was accumulated from the experimental and results sections from previous work.[16–21]

The molecular weight (mW) and solvent molecular weight (mW_s) for the compound's relative polymorphic form and the solvents used, from the experimental sections,[16–21] were calculated which proceeded with the normalization of these values to be utilized as inputs for the ANN model. Preparation of seed material and the utilization of different characterization methods for the nine organic compounds, from previous work,[16–21] allowed for generating the following data used as inputs to train and test the ANN model: mean particle size (L_0), volumetric shape factors (f_v), surface shape factors (f_s), crystallographic density (ρ_c), and molecular volume (mV). The experimental procedure, for seed preparation and the characterization methods, specific to each compound (in terms of polymorphic form and the solvent in which it's dissolved) can be described elsewhere.[16–21] To calculate the mean particle size (L_0) of each compound, a G3 Morphology particle size and shape analyser allowed for determination of the particle size distribution to evaluate the average mean particle sizes. The volumetric shape factors were computed from the slope of a linear plot involving the total mass (w) and total number (N) of crystals, their average crystal length $\bar{L}$ and crystal density (ρ_c)[22]:

$$f_v = \frac{w}{(\rho_c N(\bar{L}^3))} \tag{1}$$

The G3 Morphology particle size and shape analyser provided N and $\bar{L}$ for all particles of a sample containing a mass (w) of crystals that are determined by weighing. Similarly, surface shape factors (f_s) were determined from the slope of a linear relationship between the sum of the surface area (total surface area, S, is equal to $\sum s$) of each individual crystal in the population analysed, and the product of the number of particles times the average crystal length squared[22]:

$$f_s = \frac{S}{(N(\bar{L}^2))} \tag{2}$$

SEM images from previous work,[16–21] allowed for determination of crystal habit for each organic compound, and estimation for 2-D and 3-D surface area where an average of the 3-D/ 2-D surface area ratio is established to calculate the total 3-D surface area required for deducing area shape factors. The crystal density (ρ_c) and molecular volume (mV) was calculated using the corresponding .cif files, acquired from the CCDC database, using the software Materials Studio 7.0.[16–21]

The isothermal seeded desupersaturation (ISD) experiments performed on the nine organic compounds, from previous work,[16–21] provided information needed in retrieving the additional inputs for training and testing the ANN model, in addition to calculating the linear crystal growth rate, $G\ (m/s)$, and overall crystal growth rate, $R_g\ (kg/m^2s)$, to be utilized as outputs for the neural network. Execution of the ISD experiments permitted the generation of the following operating variables to apply as inputs for the ANN: temperature (T), agitation speed (v), relative supersaturation $(S-1)$, initial supersaturation $(S-1)_i$, lattice energy (E_l) and activation energy (E_a). For the purpose of this study, the temperature (T) and agitation speed (v), were used as inputs, once the operating variables were retrieved from the specific experimental procedures[16–21]for each of the organic compounds. The fitting of desupersaturation data from the empirical power law equation (eq 3) from previous work,[16–21] permitted the generation of each compounds' plot of supersaturation decay, (S-1), against time.

$$G = \frac{dL}{dT} = k_g(S-1)^g \tag{3}$$

These plots were inputted into the software GetData Graph Digitizer 2.26 to extrapolate the data for the initial supersaturation $(S-1)_i$ values for each of the compounds.

Fitting the power law equation (eq 4) to the experimental data, allowed for the estimation of the activation energies E_a of each compound, relative to previous work.[16–21]

$$G = k_{g0}\exp(\frac{-E_a}{RT})(S-1)^g \tag{4}$$

The computation of the compounds' lattice energy E_l , was achieved by acquiring the compounds' corresponding .cif files from the CCDC database, where calculations were executed through the software Materials Studio 7.0.

Crystal growth rate is usually specified as the rate of change of an overall linear dimension, L, of a crystal, i.e., the linear growth rate, $G\ (m/s)$ is the measurement of the fastest growing face on the crystal:

$$G = \frac{dL}{dT} = k_g(S-1)^g \tag{5}$$

From the experimental sections,[16–21] each organic compounds' plots of linear crystal growth rate, $G\ (m/s)$, against relative supersaturation $(S-1)$ were represented through the fitting of the empirical power law equation (eq 4). Referring to Eq 2 for the calculation of the volumetric shape factor, f_v, if the crystal shape is assumed to be constant, Haüys Law[23] (eq 6), in addition to the assumption of nucleation, growth rate dispersion, agglomeration and breakage to be negligible, the supersaturation decay $(S-1)$ of the ISD experiments can be calculated by the desupersaturation mass balance, with respect to the increase of size of seed crystal, $\overline{L}$:

$$\frac{w}{w_0} = (\frac{L}{L_0})^3 \tag{6}$$

$$\triangle c(t) = \triangle c_0 - [(\frac{\overline{L}}{\overline{L_0}})^3 - 1)]\frac{W_0}{M} \tag{7}$$

with the variation of crystal size given by:

$$\overline{L} = \overline{L}_0 + \int_0^t G((S-1)(t))dt \tag{8}$$

Similarly, these plots of $G\ (m/s)$ against $(S-1)$ were opened in GetData Graph Digitizer 2.26 to extrapolate the relative supersaturation $(S-1)$, in addition to the linear crystal growth rate, G (m/s) (eq 4). The collective data for all inputs used for the ANN

model are referred to in Table 1.

For interpreting the linear crystal growth rate, $G\ (m/s)$, in respect of the mass deposition rate, the calculated G values can be computed in terms of the overall crystal growth rate, $R_g\ (kg/m^2s)$, described as the rate of mass deposition that is accumulated in a meter squared per second throughout the crystal:

$$R_g = \frac{3 f_v \rho_c}{f_s} G \tag{9}$$

Performing recalculations on all linear crystal growth rate values, $G\ (m/s)$, in terms of Eq 6, generated all output data, specific to the overall crystal growth rate, $R_g\ (kg/m^2s)$. The linear, $G\ (m/s)$, and overall, $R_g\ (kg/m^2s)$, crystal growth rate values were used as outputs for training and testing the ANN model, compiled in Table 2.

3 Multiple Non-Linear Regression Analysis

Prediction of crystal growth kinetics is traditionally performed by modelling the operating variables that affect crystallization kinetics with a semi-empirical expression that is determined through the regression of experimental data.[12,13] This traditional approach for kinetic predictions have evidently shown high variance in the mean relative error difference between the experimental and predicted data. In Yang and Wei's study to predict the crystal growth kinetics of ciprofloxacin, an attempt to predict G values through the regression of an empirical expression concluded with a mean relative error, between the experimental and predicted data, to be as high as 19.5%.[13] Similarly, the crystal growth kinetics of sucrose was analysed using a regressed semi-empirical expression with a correlation coefficient between the experimentally determined R_g values of sucrose and the semi-empirical expression to be 0.748.[12] Prediction of crystal growth kinetics using regressed semi-empirical expressions have resulted in inaccuracies due to limitations arising from the inability to generalize an expression correlating all variables involved in a specific crystallization process. In this study, an attempt to correlate the growth rate data from the desupersaturation experiments,[16–21] with the input variables from each of the organic compounds (in Table 1), was proceeded by using multiple nonlinear regression (MNLR). This form of regression analysis involves modelling a mathematical function consisting of a nonlinear combination of the model's parameters, dependant on one or more independent variables. These mathematical functions are generally influenced on the basis of empirical expressions from a related process to the experimental data. MNLR involves an iterative process which is applicable to computer operations such as Microsoft Excel, MATLAB, and Python. In this case, MNLR analysis was utilized for the determination of a semi-empirical expressions that correlated the linear crystal growth rate, $G\ (m/s)$, with the input and output data from the constructed database.

The prediction of nonlinear combinations of the model's parameters was executed by minimizing the error distribution between the linear crystal growth rate data, $G_{Experimental}$, from the constructed database, and the determined linear crystal growth rate data, G_{MNLR}, from the regressed semi-empirical expression. This was performed by minimizing the value of the sum of the squared residuals (SSR) :

$$SSR = \sum (G_{MNLR} - G_{Experimental})^2 \tag{10}$$

The iterative process to perform the minimization of the error distribution was achieved by using the Solver add-in with the Microsoft Excel Application. Minimizing the SSR value results in an inverse correlation of the coefficient of determination (r^2) for G_{MNLR}, that can be defined as:

$$r^2 = \frac{(G_{MNLR} - \overline{G_{Experimental}})^2}{\sum[(G_{MNLR} - \overline{G_{Experimental}})^2 + (G_{MNLR} - \overline{G_{Experimental}})^2)]} \tag{11}$$

To initialize the MNLR model, approximation of the model's parameters was executed by correlating the individual operating variables against the linear crystal growth rate data, $G_{Experimental}$. The predicted semi-empirical expression from the MNLR analysis technique is described in Eq 12:

$$G_{MNLR} = 3.5748 \cdot 10^{22} \times L_0^{2.6199} \times T^{16.541} \times v^{2.1337} \times (S-1)^{0.6465} \times (S-1)_i^{0.1623} \times mV^{2.2889} \times mW^{2.5524} \times mW_s^{1.4252} \times E_l^{1.9957} \tag{12}$$

The data used for prediction of eq 12 was from the training data set for the ANN model. The minimal value for the sum of the squared residuals (SSR) equated to a high value of 188075.244, indicating that the MNLR model is poorly representative of the training data set. The coefficient of determination (r^2) was measured as 12.48%, demonstrating the MNLR model's inaccuracy to predict the linear crystal growth rate of a combination of multiple organic compounds simultaneously. Figure 2 displays the plot of linear and overall growth rates determined from experiments, $G_{Experimental}$ and $R_{g,Experimental}$, in addition to the prediction from the MNLR analysis, G_{MNLR} and $R_{g,MNLR}$. Due to substantial inaccuracies of predicted G_{MNLR} values, the recalculated $R_{g,MNLR}$ values determined from eq 9 proceeded with a similar outcome of the SSR and r^2 values equating to 6.852 and 11.40%. To improve the correlations between the operating variables and the crystal growth rate data, an ANN model was created with the data from the constructed database, to train and test the neural network for increased accuracy of crystal growth rates, G and R_g, of multiple compounds.

4 Neural Network Modelling

Neural networks are systems generally consisting of a large number of simple processing units, i.e. neurons, that are capable of storing experiential data with the ability to use this data for solving complex and nonlinear problems. Neural networks can be compared to the brain in a sense that the knowledge or data is acquired by the ANN through a specific learning process, in addition to the ANN compiling of synaptic weights, i.e. interneuron connection strengths, that are used to store the acquired data.[24] In this study, the ANN model constructed consists of a multiple layer feed-forward back propagation neural network with 13 elements in the input layer, one hidden layer, and two elements in the output layer. The construction of this ANN model was utilized to accurately simulate the linear and overall crystal growth rates,

Table 1 Operating Variables Used To Train And Test The ANN Model

Compounds	L_0 [μm]	T [K]	v [rpm]	$(S-1)$ [-]	$(S-1)_i$ [-]	mV [Å^3]	mW [g/mol]
AAP Form I (in MeOH)	125	288 - 303	250	0.0015 - 0.1202	0.1135 - 0.1204	188.49	151.17
CBMZ Form III (in MeOH)	160	288 - 303	250	0.0107 - 0.1724	0.1586 - 0.1826	285.67	236.27
CUR Form I (in IPrOH)	67.8	288 - 318	200	0.01327 - 1.0541	1.0085 - 1.0667	338.7	368.40
FF Form I (in MeOH)	177	288 - 303	250	0.0057 - 0.1756	0.1259 - 0.1135	458.85	360.83
PCM Form II (in EtKOH)	127	283 - 308	190	0.0002 - 0.1186	0.0429 - 0.1186	134.24	142.16
PCM Form II (in IPrOH)	127	283 - 308	190	0.0010 - 0.1044	0.0573 - 0.1044	134.24	142.16
PCM Form III (in EtKOH)	145	283 - 308	190	0.0004 - 0.1702	0.1455 - 0.1705	136.27	142.16
PCM Form III (in IPrOH)	145	283 - 308	190	0.0008 - 0.1710	0.1432 - 0.1712	136.27	142.16
PCM Form III (in MeOH)	119	288 - 303	250	0.0013 - 0.0836	0.0331 - 0.0838	136.27	142.16
RIS Form A (in MeOH)	123	288 - 303	250	0.0124 - 0.2752	0.1979 - 0.2521	520.01	410.49
SM Form I (in AC)	328	283.15 - 298.15	500	0.0002 - 0.0304	0.0098 - 0.0284	171.00	137.14
SM Form I (in MeCN)	328	283.15 - 298.15	500	0.0002 - 0.0301	0.0094 - 0.0285	171.00	137.14
SM Form I (in EA)	328	283.15 - 298.15	500	0.0007 - 0.0558	0.0216 - 0.0559	171.00	137.14
SM Form I (in MeOH)	328	283.15 - 298.15	500	0.0002 - 0.0330	0.0111 - 0.0330	171.00	137.14
SA (in AC)	187.5	283.15 - 298.15	180	0.0006 - 0.0323	0.0124 - 0.0323	159.00	138.12
SA (in MeCN)	187.5	283.15 - 298.15	180	0.0004 - 0.0360	0.0118 - 0.0360	159.00	138.12
SA (in EA)	187.5	283.15 - 298.15	180	0.0001 - 0.0310	0.0091 - 0.0310	159.00	138.12
SA (in MeOH)	187.5	283.15 - 298.15	180	0.0002 - 0.0436	0.0297 - 0.0436	159.00	138.12
TBM Form I^L (in MeCN)	215	268 - 280	270	0.0016 - 0.1837	0.0909 - 0.1834	242.56	270.35
TBM Form I^L (in EtKOH)	215	268 - 280	270	0.0009 - 0.2127	0.1444 - 0.2176	242.56	270.35
TBM Form I^L (in EA)	215	268 - 280	270	0.0027 - 0.3243	0.0971 - 0.3242	242.56	270.35
TBM Form I^L (in IPrOH)	215	268 - 283	270	0.0012 - 0.1898	0.0717 - 0.1895	242.56	270.35
TBM Form I^L (in TOL)	215	277 - 283	270	0.0073 - 0.8414	0.1529 - 0.8430	242.56	270.35

Compounds	ρ_c [kg/m^3]	f_v [-]	f_s [-]	mW_s [g/mol]	E_l [kJ/mol]	E_a [kJ/mol]
AAP Form I (in MeOH)	1298	0.303	1.967	32.042	-515.24	47.35
CBMZ Form III (in MeOH)	1347	0.164	2.110	32.042	-129.09	38.00
CUR Form I (in IPrOH)	1390	0.460	3.142	60.100	-1014.19	37.66
FF Form I (in MeOH)	1285	0.215	1.696	32.042	-325.33	36.29
PCM Form II (in EtKOH)	1355	0.146	2.051	46.069	-267.24	39.28
PCM Form II (in IPrOH)	1355	0.146	2.051	60.096	-267.24	58.40
PCM Form III (in EtKOH)	1366	0.224	2.232	46.069	-534.08	65.08
PCM Form III (in IPrOH)	1366	0.224	2.232	60.096	-534.08	62.25
PCM Form III (in MeOH)	1366	0.224	2.232	32.042	-534.08	45.20
RIS Form A (in MeOH)	1291	0.274	2.344	32.042	-760.60	38.21
SM Form I (in AC)	1330	0.01	0.42	58.08	-781.57	58.50
SM Form I (in MeCN)	1330	0.01	0.42	41.05	-781.57	91.20
SM Form I (in EA)	1330	0.867	5.384	88.11	-781.57	80.70
SM Form I (in MeOH)	1330	0.01	0.42	32.042	-781.57	88.40
SA (in AC)	1440	0.01	0.42	58.08	-350.13	44.60
SA (in MeCN)	1440	0.01	0.42	41.05	-350.13	22.10
SA (in EA)	1440	0.01	0.42	88.11	-350.13	67.20
SA (in MeOH)	1440	0.01	0.42	32.042	-350.13	40.10
TBM Form I^L (in MeCN)	1252	0.867	5.384	41.053	-722.68	37.50
TBM Form I^L (in EtKOH)	1252	0.867	5.384	46.069	-722.68	63.00
TBM Form I^L (in EA)	1252	0.867	5.384	88.106	-722.68	72.70
TBM Form I^L (in IPrOH)	1252	0.867	5.384	60.096	-722.68	45.20
TBM Form I^L (in TOL)	1252	0.867	5.384	92.141	-722.68	83.10

G_{ANN} and $R_{g,ANN}$, for multiple organic compounds collectively. The feed-forward aspect of a multiple layer neural network only specifically allows the input data to flow directionally towards the neurons within the individual layers to form an output with relatively accurate approximation of any function for the given input conditions. ANN models are useful for predicting complex relationships, such as crystallization kinetics, due to the non linearity and adaptive properties of a neuron allowing for computation of any linear or nonlinear relationship. Neural networks have ability to adapt to surrounding environments by adjusting their layer weights to deal with minor changes in the operating conditions of a data set. Particularly, neural networks with multiple layers have a tendency of adaptivity due to interpretation of error deviations from resultant vectors in the hidden layers of the network.[25]

A detailed structure of the constructed ANN model, depicting the multiple layers and the flow of information throughout the neural network, is illustrated in Figure 3. Matrix multiplication can explicitly describe the process of feed-forward within a neural network. In this network, the 13 elements in the input layer, comprising of 3,251 input vectors, p, each, are emitted and connected to each neuron input through a weight matrix, W. Eq 13 depicts the weight matrix for the ANN model, specific to the input-hidden layer:

$$W_{I\to H} = \begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,42263} \\ \vdots & \vdots & \vdots & \vdots \\ w_{10,1} & w_{10,2} & \cdots & w_{10,42263} \end{bmatrix} \tag{13}$$

Every neuron calculates the weighted sum of the inputs, with addition to a bias, b, described in eq 14, to produce an output for the hidden layer, once it is passed through an activation function. These calculations, from eq 14, illustrate a correlation of the matrix product of weight matrix, W, inclusive of a bias, times

Table 2 Outputs To Train And Test The ANN Model

Compounds	$G\ [m/s]\ \times 10^7$	$R_g\ [kg/m^2s]\ \times 10^7$
AAP Form I (in MeOH)	3.57×10^{-3} - 1.17	2.14 - 702.67
CBMZ Form III (in MeOH)	0.01 - 0.31	3.17 - 96.19
CUR Form I (in IPrOH)	2.98×10^{-3} - 1.90	1.82 - 1157.85
FF Form I (in MeOH)	0.01 - 3.73	5.49 - 1822.85
PCM Form II (in EtKOH)	2.03×10^{-3} - 1.91	0.59 - 551.47
PCM Form II (in IPrOH)	1.85×10^{-3} - 0.66	0.54 - 189.90
PCM Form III (in EtKOH)	3.83×10^{-3} - 2.01	1.58 - 825.49
PCM Form III (in IPrOH)	1.95×10^{-3} - 0.53	0.80 - 217.77
PCM Form III (in MeOH)	9.04×10^{-3} - 2.99	3.72 - 1232.15
RIS Form A (in MeOH)	3.00×10^{-3} - 0.92	1.36 - 417.38
SM Form I (in AC)	5.7×10^{-3} - 1.68	0.54 - 160.07
SM Form I (in MeCN)	9.64×10^{-3} - 1.60	0.92 - 152.34
SM Form I (in EA)	0.01 - 2.22	6.85 - 1429.20
SM Form I (in MeOH)	0.01 - 3.77	1.41 - 358.01
SA (in AC)	0.02 - 3.20	1.20 - 329.57
SA (in MeCN)	0.01 - 1.59	1.33 - 163.81
SA (in EA)	4.28×10^{-3} - 1.69	4.40 - 1741.968
SA (in MeOH)	8.39×10^{-3} - 3.49	0.86 - 358.57
TBM Form I^L (in MeCN)	0.53 - 97.11	317.50 - 58737.65
TBM Form I^L (in EtKOH)	0.13 - 45.42	78.96 - 27469.54
TBM Form I^L (in EA)	0.06 - 63.74	37.24 - 38552.91
TBM Form I^L (in IPrOH)	0.14 - 36.38	85.99 - 22006.76
TBM Form I^L (in TOL)	0.06 - 9.53	34.46 - 5765.02

a column of inputs with an additional value of 1, as it passes through an activation function, *tansig*, deciphered by the neural network. Eq 14 can be condensed into a formula describing the outputs emitted into the hidden layer, seen in eq 15, where H is the hidden layer's outputs, W is the weight matrix, p is the input vectors, and b is the hidden layer's bias:

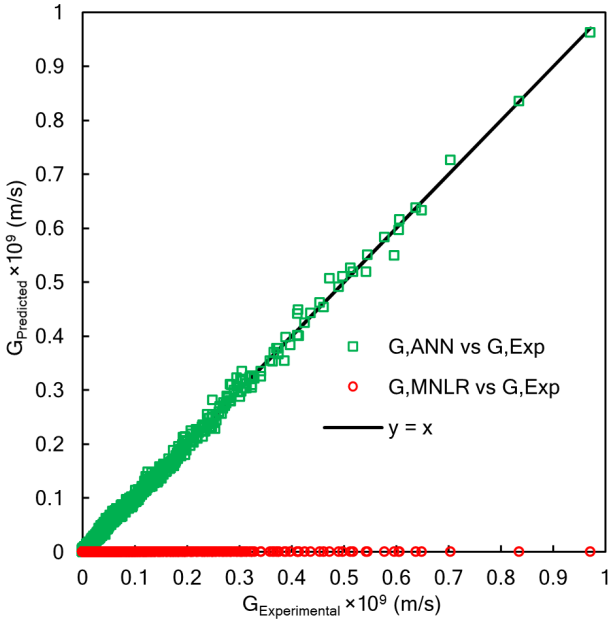
$$H = tansig\left(\begin{bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,42263} & b_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{10,1} & w_{10,2} & \cdots & w_{10,42263} & b_{10} \end{bmatrix} \cdot \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{42263} \\ 1 \end{bmatrix}\right)$$

$$= tansig\begin{bmatrix} (w_{1,1}\times p_1)+(w_{1,2}\times p_2)+\cdots+(w_{1,42263}\times p_{42263})+b_1 \\ \vdots \\ (w_{10,1}\times p_1)+(w_{10,2}\times p_2)+\cdots+(w_{10,42263}\times p_{42263})+b_{10} \end{bmatrix} \quad (14$$

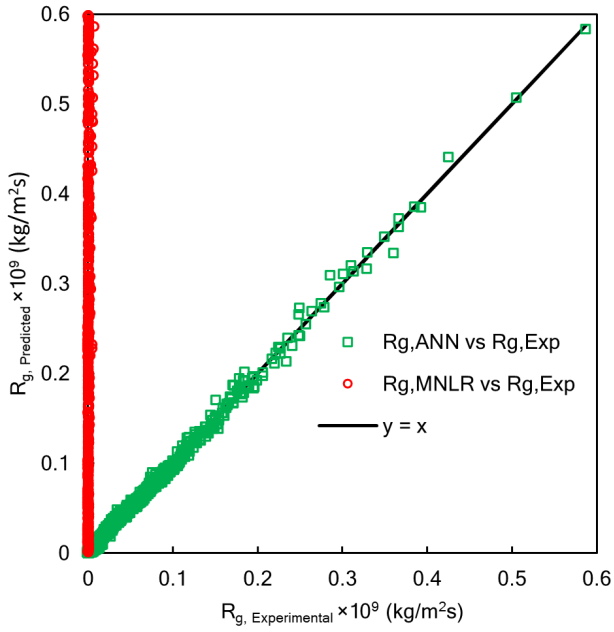
$$H = tansig((W\times p)+b) \quad (15)$$

The hidden layer's product, H, feeds forward through the ANN model and emits into the output layer as an input into the output's neurons. Similarly, this process repeats, instead with a *purelin* activation function, that returns a product, O, to the output layer. Thus, the O outputs are the values that predict G and R_g for the training data set.

Figure 4 denotes the training strategy of the ANN model, where the input vectors and the corresponding output vectors were utilized for training the neural network until accurate approximation of the activation functions were determined. It was noted that the use of a tansigmoid function in the hidden layer, and a linear function in the output layer, permitted effective approximation of crystal growth kinetics. In this study, the input vectors



(a)



(b)

Fig. 2 Parity plot of (a) experimentally obtained linear crystal growth rate, $G_{Experimental}$, against the predicted linear crystal growth rates by MNLR, G_{MNLR}, and ANN, G_{ANN}, analysis, and (b) experimentally obtained overall crystal growth rate, $R_{g,Experimental}$, against the predicted linear crystal growth rates by MNLR, $R_{g,MNLR}$, and ANN, $R_{g,ANN}$, analysis on the training data set.

were as follows: L_0, T, v, $(S-1)$, $(S-1)_i$, mV, mW, ρ_c, f_v, f_s, mW_s, E_l and E_a. The linear and overall crystal growth rates, G and R_g, were defined as the output vectors.

Simulation of the linear and overall crystal growth rates were performed using the Deep Learning Toolbox Version 14.1 from

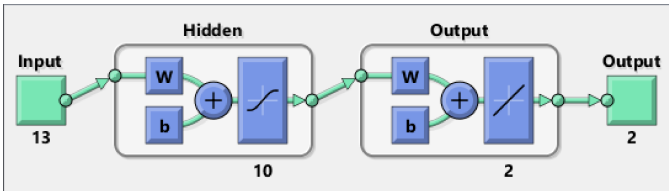


Fig. 3 A detailed structure of the constructed ANN model illustrating the flow of data within the neural network.

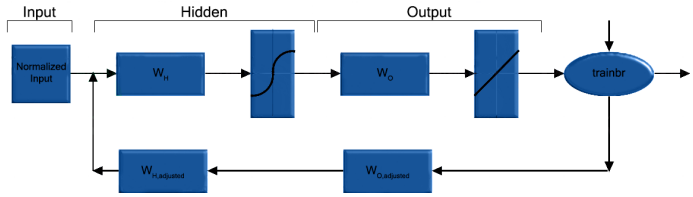


Fig. 4 Training strategy utilized by bayesian regularization algorithm to train the ANN model.

MATLAB,[15] where the operating variables, in Table 1, and the determined crystal growth rates, in Table 2, were set as the input and target vectors for the neural network. Normalization of data is critical in increasing the efficiency of the learning process. Specifically, the input and target vectors were normalized prior to training, falling in the interval of -1 and +1, to calibrate with the Bayesian algorithm and the *tansig* function, in addition to maintaining a mean close to zero. The tansigmoid activation function was used over the sigmoid function in the neural network due to the ability to contain both positive and negative values, which increases the flexibility for the neurons to learn. Conversely, utilizing a sigmoid function within a neural network results in the emission of outputs of only positive values, leading to increasingly stifled learning capacity.

There are several different algorithms that can be used in the training process of an ANN. The Levenberg-Marquadt, *trainlm*, algorithm deemed unfavourable for this neural network compared to the Bayesian framework, *trainbr*, due to the extensive number of weights applied in the ANN model. *Trainlm* tends to have the fastest convergence on function approximation problems, but the advantage of using this algorithm dramatically decreases as the number of weights in the network increases. One of the problems during the training process in a neural network is overfitting. The occurrence of overfitting increases when an excessive number of neurons are used in the hidden layer. The determination of overfitting occurring in a network can be devised by high levels of error deviations on the testing data set between the experimentally determined crystal growth rates with the determined values from the ANN model. To prevent overfitting within a neural network, numerous attempts of using different numbers of neurons in the hidden layer were trialed on the training data set to ensure the optimal ANN size was determined, leading to increased performance of prediction of the linear and overall crystal growth rates. Overtraining is another implication that reduces the capacity to optimize the propagation function during the training process. This occurs when the training time of the neural network is prolonged, resulting in memorization of the training data which causes the performance of the ANN to reduce due to the lack of generalization of the experimental data. Thus, to avoid such complications, regularization of the data can be computed by using the Bayesian algorithm. In Figure 4, a visualization of the regularization process is evident where the performance of the transfer function is modified, in addition to noise reduction, which is performed by using the Bayesian framework of David MacKay[26] as the training strategy of the neural network due to the automatic determination of the optimal regularization parameters.

Within the Bayesian framework, there is an assumption of the weights and biases to be random variables with specified distribution in the neural network. Statistical techniques can be used to estimate the regularization parameters as there is a correlation between the unknown variance associated with these distributions and the regularization parameters. The function *trainbr* implements the Bayesian algorithm in the ANN where scaling of the input and target vectors between -1 and 1 deems to be increasingly effective for the algorithm to perform competently. Optimum network size of the ANN model is determined by the Bayesian algorithm on account of the ability of the Bayesian framework to measure the effective weights and biases, in addition to determination of the number of effective parameters, used in the neural network.

Details of the ANN model and the network size is described in Table 3, with a 13-10-2 network size. 10 neurons in the hidden layer was determined to be suitable for the ANN model for representing the crystal growth rate data. The architecture of the neural network is typically deduced by a trial-and-error basis where consideration of the number of neurons in the hidden layer can be estimated by non-empirical assumptions for an initial prediction of hidden layer neurons. It was noted that increasing the number of hidden layer neurons greater than 10 resulted in no affect on the performance of the ANN model for prediction of crystal growth rates.

Table 3 Information of the ANN Model Used in Predicting G and R_g values

Parameters	Description
Input Layer	13 Neurons
Hidden Layer	10 Neurons
Output Layer	2 Neurons
Data Points Used for Training ANN	48,765
Data Points Used for Testing ANN	12,195
Function Used in Hidden Layer	Tansigmoid Function
Function Used in Output Layer	Linear Function

For ANN models with a size greater than a 13-10-2 network, the Bayesian algorithm automatically chooses the weights and biases resulting in the number of weights and biases to be constant within consideration of the increased network size. In this study, 133 network parameters were used to optimize the 48,765 data points for simulation of the linear and overall crystal growth kinetics of the multiple organic compounds. The number of degrees of freedom of the ANN model are typically independent from the

Bayesian framework, resulting in a constant number of effective parameters in the neural network.

For prediction of crystal growth kinetics, a network of 48,765/133 degrees of freedom was adequate in accurately simulating the training and test data sets. Convergence of the constructed ANN model occurred when there was a constant correlation between the sum of the squared error with the sum of the squared weights, following with optimal validation of the data proceeding once the minimal mean squared error was achieved at 844 epochs.

Figure 2 illustrates a parity plot of (a) predicted G values, and (b) predicted R_g values, by the ANN model, against the experimental G and R_g values, specific to the training data set. It is evident that the constructed ANN model depicts the crystal growth kinetics of multiple compounds to a highly distinguished accuracy. This level of accuracy from the ANN model was further verified from the calculated statistics of the training data set in Table 4. The coefficient of determination (eq 11), r^2, can be broken down into segments to understand the correlation between the predicted and experimental crystal growth rate data. The ratio between the sum of the squared average residuals, $SSAR$, and the sum of the squared total, SST, determines the value calculated for the coefficient of determination, r^2:

$$SSAR = (G_{ANN} - G_{Exp\,Average})^2 \; or \; (R_{g,ANN} - R_{g,Exp\,Average})^2 \quad (16)$$

$$SST = SSAR + SSR \quad (17)$$

The r^2 values determined between the experimental, $G_{Experimental}$ and $R_{g,Experimental}$, values and the predicted, G_{ANN} and $R_{g,ANN}$, values was measured as 0.9970 and 0.9958 ($\sim$1). These high r^2 values imply that the constructed ANN model is exceptionally accurate in the prediction of both linear and overall crystal growth kinetics of the multiple compounds in the data set. Conversely, the conventional MNLR analysis was considered to be an unsuitable model for crystal growth rate predictions, demonstrated by a low degree of r^2 values, 0.1248 and 0.1144.

For validation of the trained ANN model, new inputs that were distinct from the training data set, were fed into the neural network to test the performance of the predicted crystal growth rates. Figure 5 and 6 shows the plot of predicted G and R_g values by the ANN model in relation to the experimentally determined G and R_g values from the test data set. Similarly to the training data set, the ANN model excelled in precisely predicting the crystal growth rates of the multiple organic compounds. Table 5 demonstrates this high level of accuracy from the ANN model's predictions with high correlation coefficients equating to r^2 values of 0.9959 and 0.9945.

Table 6 displays the average percentage deviations of each organic compound between the experimentally determined G and R_g values from the trained neural network for the new test data set. The average percentage deviations for the determined linear crystal growth rates ranged from relatively low values of 0.0598% to 3.3247%. Higher average percentage deviations from each of the organic compounds occurred with the predicted overall crystal growth rates ranging from 0.0834% to as high as 20.1383%.

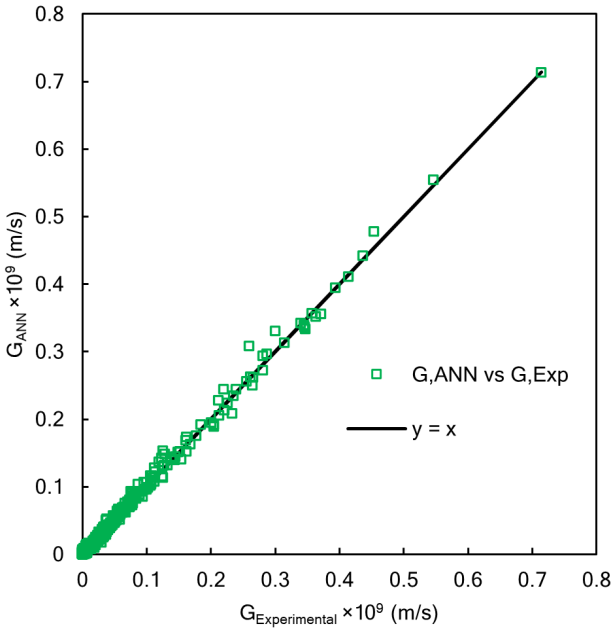


Fig. 5 Parity plot of experimentally obtained linear crystal growth rate, $G_{Experimental}$, against the predicted linear crystal growth rates by MNLR, G_{MNLR}, and ANN, G_{ANN}, analysis on the test data set.

Lack of calculated shape factor data on specific compounds resulted in alternatively using approximate shape factors from idealized crystal shapes[27] on the basis of the crystal lattices' SEM images. Consequently, there is an indication that utilizing approximate shape factors for specific compounds of the neural network resulted in abnormal R_g values that were calculated, in turn increasing the percentage deviations for compounds such as Salicylamide. Another potential factor that could have skewed the prediction of crystal growth kinetics in the neural network is unusual crystal behaviour of specific polymorphic forms within a solvent during the desupersaturation experiments that altered the experimentally determined G and R_g values. However, the developed neural network model had incremental capability of generalization in predicting linear and overall crystal growth kinetics of multiple compounds simultaneously.

The crystallization kinetics of organic compounds is a process with incremental complexity that involves extensive operating variables that affect the crystal growth rate of multiple compounds. Additionally, proposing a generalized expression to correlate operating variables between the multiple organic compounds for the determination of crystal growth kinetics is a tremendously intricate task. In this study, the trained ANN model determined highly accurate estimations in respect of predicting linear and overall crystal growth rates of multiple organic compounds simultaneously. The ability for the ANN model to generalize in terms of the constructed database emphasised on the unprecedented advantage neural networks possess over conventional methods such as analysis through multiple non-linear regression.

Table 4 Statistical Analysis of Training Data for G_{MNLR} vs $G_{Experimental}$, G_{ANN} vs $G_{Experimental}$, $R_{g,MNLR}$ vs $R_{g,Experimental}$ and $R_{g,ANN}$ vs $R_{g,Experimental}$

Statistics	G_{MNLR} vs $G_{Experimental}$	G_{ANN} vs $G_{Experimental}$	$R_{g,MNLR}$ vs $R_{g,Experimental}$	$R_{g,ANN}$ vs $R_{g,Experimental}$
Exp Average	2.8726	2.8726	0.0165	0.0165
SSAR	26826.2269	160757.0486	0.8851	5.9415
SSR	188075.2440	490.9424	6.8520	0.0253
SST	214901.4710	161247.9910	7.7371	5.9668
r^2	0.1248	0.9970	0.1144	0.9958

Table 5 Statistical Analysis of Test Data Set for G_{ANN} vs $G_{Experimental}$ and $R_{g,ANN}$ vs $R_{g,Experimental}$

Statistics	G_{ANN} vs $G_{Experimental}$	$R_{g,ANN}$ vs $R_{g,Experimental}$
Exp Average	0.0301	0.0172
SSAR	4.0430	1.4949
SSR	0.0165	0.0082
SST	4.0595	1.5031
r^2	0.9959	0.9945

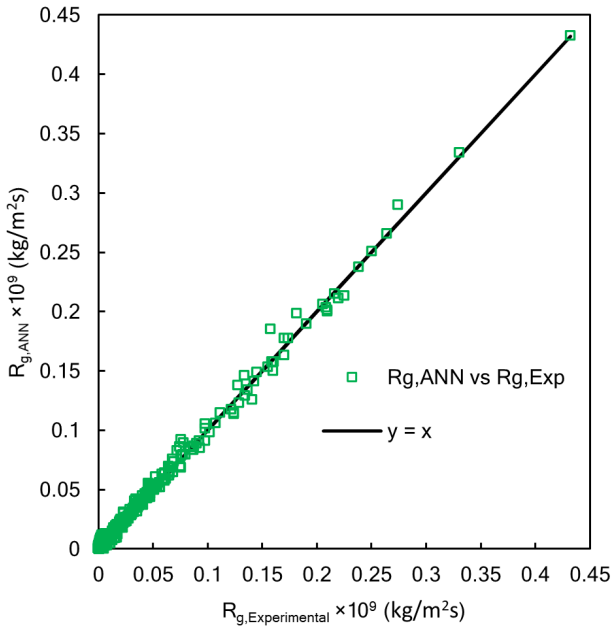


Fig. 6 Parity plot of experimentally obtained overall crystal growth rate, $R_{g,Experimental}$, against the predicted linear crystal growth rates by MNLR, $R_{g,MNLR}$, and ANN, $R_{g,ANN}$, analysis on the test data set.

5 Conclusion

The crystal growth kinetics from different polymorphic forms of Acetaminophen, Carbamazepine, Curcumin, Fenofibrate, Piracetam, Risperidone, Salicylamide, Salicylic Acid and Tolbutamide, in a range of solvents, were predicted with the application of an artificial neural network (ANN) model. Constructing a feed-forward back propagation neural network, consisting of a tan-sigmoid activation function in the hidden layer with a linear propagation function in the output layer, in addition to the implementation of the Bayesian algorithm, permitted a viable model that was trained to model the linear and overall crystal growth kinetics of multiple compounds simultaneously. The ANN model was trained with the constructed database, consisting of multiple operating variables, proceeding with the determination of the neural network to be effective in prediction of the linear and overall crystal growth rates of the multiple organic compounds.The coefficient of determination, in respect of the training data set, to represent the experimental $G_{Experimental}$ and $R_{g,Experimental}$ values with the predicted G_{ANN} and $R_{g,ANN}$ values by the trained neural network, was specified as 0.9970 and 0.9958 (~ 1). A new test data set was implemented to the trained ANN model resulting in accurate simulation of the linear and overall crystal growth kinetics for the multiple organic compounds. Under these conditions, the coefficient of determination (r^2) between the experimentally determined $G_{Experimental}$ and $R_{g,Experimental}$ values with the predicted G_{ANN} and $R_{g,ANN}$ values by the constructed ANN model, was observed to be 0.9959 and 0.9945. The ANN model had exceptional ability in representing the crystal growth kinetics from different polymorphic forms of multiple organic compounds dissolved in a range of solvents, deeming it to be eminently precise which validated the neural network model's viability. Performance of a conventional multiple nonlinear regression (MNLR) analysis was compared against the ANN model, for the training data set, leading to a poorly representative model with significant inaccuracies in predicting the crystal growth kinetics due to the inability to correlate the operating variables with the crystal growth rate data.

Table 6 Average Percentage Deviation Between Each Organic Compound's G_{ANN} vs $G_{Experimental}$ and $R_{g,ANN}$ vs $R_{g,Experimental}$

Compounds	G Deviation (%)	R_g Deviation (%)
AAP Form I (in MeOH)	0.6699	0.6639
CBMZ Form III (in MeOH)	0.5005	0.5942
CUR Form I (in IPrOH)	1.0226	0.9267
FF Form I (in MeOH)	1.0544	3.2875
PCM Form II (in EtKOH)	0.7610	3.9053
PCM Form II (in IPrOH)	0.8825	11.8676
PCM Form III (in EtKOH)	1.0696	1.2313
PCM Form III (in IPrOH)	0.6148	4.2934
PCM Form III (in MeOH)	1.5668	2.6086
RIS Form A (in MeOH)	0.4318	1.9989
SM Form I (in AC)	0.3392	6.0885
SM Form I (in MeCN)	0.7942	20.1383
SM Form I (in EA)	0.18850	0.3762
SM Form I (in MeOH)	0.1965	3.0669
SA (in AC)	1.4952	11.2023
SA (in MeCN)	0.6029	7.2388
SA (in EA)	3.3247	2.3246
SA (in MeOH)	0.4956	7.5851
TBM Form I^L (in MeCN)	0.1366	0.1287
TBM Form I^L (in EtKOH)	0.1008	0.0834
TBM Form I^L (in EA)	0.0880	0.1183
TBM Form I^L (in IPrOH)	0.0914	0.0915
TBM Form I^L (in TOL)	0.0598	0.1042

Supporting Information Available: The Matlab file of the constructed ANN is available in the .m file attached (ANN.m), with the data sets compiled in the .mat file (ANNWorkspace.mat).

References

1 *Polymorphism*, John Wiley Sons, Ltd, 2009, ch. 3, pp. 49–76.

2 A. Jones, *Crystallization Process Systems*, Butterworth-Heinemann, Oxford, 2002, pp. 123 – 154.

3 V. I. Kalikmanov, in *Classical Nucleation Theory*, Springer Netherlands, 2012, p. 17–41.

4 J. Ulrich, *Crystal Growth & Design*, 2001, **1**, 88–125.

5 S. Araokar, *arXiv preprint cs/0505016*, 2005.

6 P. D. Cristea, *GeoSpatial Visual Analytics*, Springer, 2009, pp. 59–71.

7 T. Kimoto, K. Asakawa, M. Yoda and M. Takeoka, 1990 IJCNN international joint conference on neural networks, 1990, pp. 1–6.

8 R. S. Barton, *PhD thesis*, University of Texas at Austin, 1997.

9 J. C. MacMurray and D. Himmelblau, *Computers & chemical engineering*, 1995, **19**, 1077–1088.

10 A. D. Naik and S. S. Bhagwat, *Journal of Chemical & Engineering Data*, 2005, **50**, 460–467.

11 K. V. Kumar and K. Porkodi, *Chemical Engineering Journal*, 2009, **148**, 20–25.

12 K. Vasanth Kumar, P. Martins and F. Rocha, *Industrial & engineering chemistry research*, 2008, **47**, 4917–4923.

13 M. Yang and H. Wei, *Industrial & engineering chemistry research*, 2006, **45**, 70–75.

14 C. Damour, M. Benne, B. Grondin-Perez and J.-P. Chabriat, *Journal of Food Engineering*, 2010, **99**, 225–231.

15 N. N. Toolbox, *Inc., Natick, MA*, 2004.

16 R. Soto, V. Verma, A. Lynch, B. K. Hodnett and Å. C. Rasmuson, *Crystal Growth & Design*, 2020, **20**, 7626–7639.

17 C. Heffernan, R. Soto, B. K. Hodnett and Å. C. Rasmuson, *CrystEngComm*, 2020.

18 R. Soto and Å. C. Rasmuson, *Crystal Growth & Design*, 2019, **19**, 4273–4286.

19 A. Lynch, L. Jia, M. Svard and Å. C. Rasmuson, *Crystal Growth & Design*, 2018, **18**, 7305–7315.

20 L. Jia, M. Svard and Å. C. Rasmuson, *Crystal Growth & Design*, 2017, **17**, 2964–2974.

21 R. Soto, V. Verma and Å. C. Rasmuson, *Crystal Growth & Design*, 2020, **20**, 1985–1996.

22 J. Garside, J. W. Mullin and S. N. Das, *Industrial & Engineering Chemistry Process Design and Development*, 1973, **12**, 369–371.

23 M. H. N. Story-Maskelyne and N. Story-Maskelyne, *Crystallography: A Treatise on the Morphology of Crystals*, Clarendon Press, 1895.

24 S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, 1999.

25 B. Widrow and M. A. Lehr, *International Journal of Intelligent Systems*, 1993, **8**, 453–507.

26 D. J. MacKay, *Neural computation*, 1992, **4**, 448–472.

27 J. Garside, A. Mersmann, J. Nỳvlt, I. of Chemical Engineers (Great Britain) and E. F. of Chemical Engineering. Working Party on Crystallization, *Measurement of Crystal Growth and Nucleation Rates*, Institution of Chemical Engineers, 2002.