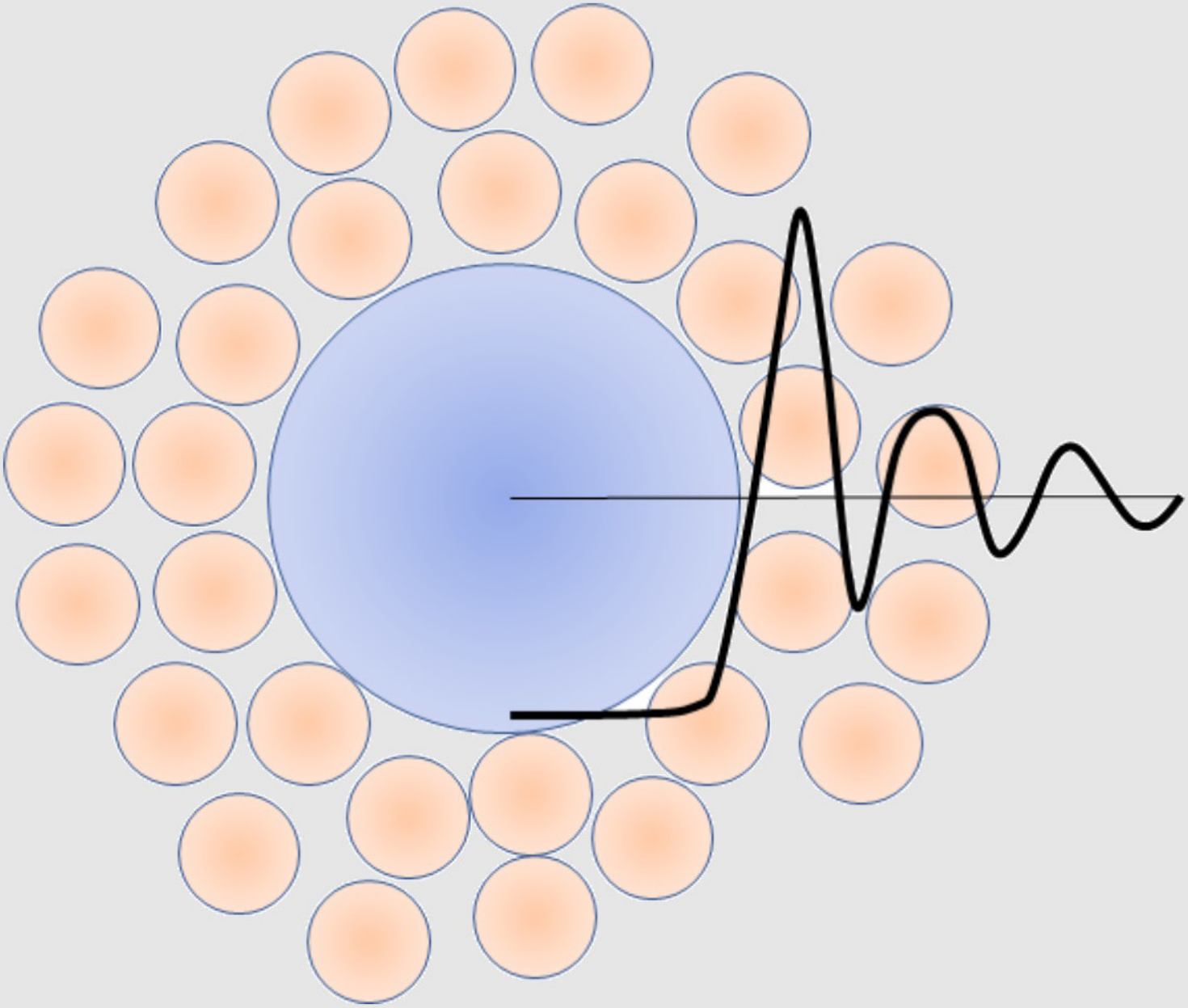
Solubility Science

Principles & Practice



Prof **Steven Abbott**

Solubility Science: Principles and Practice

Prof Steven Abbott

Steven Abbott TCNF Ltd, Ipswich, UK
and Visiting Professor, University of Leeds, UK

steven@stevenabbott.co.uk
www.stevenabbott.co.uk

Version history:

First words written	15 Sep 2015
Version 1.0.0	15 Sep 2017
Version 1.0.1.1	05 Oct 2017 (Some additions on HSP)
Version 1.0.1.2	05 Jan 2018 (An update on RDF calculations for large, complex solute molecules such as proteins)
Version 1.0.1.3	05 April 2020 (A number of typos fixed thanks to the kind checking by Lee McManus)
Version 1.0.2.0	March 2021 (Crystallization chapter added)
Version 1.0.2.1	March 2021 (Minor crystallization updates)
Version 1.0.2.1a	March 2021 (Correction on NOESY)

Contents

Preface

Those of us who have to formulate solutions for industrial applications can adopt two extreme positions with respect to how we go about it. The first way is to rely on intuition and experience, "how it's always been done". The second is to dig deep into profound theory in the hope that the answers will be found there.

My experience with formulators around the world is that the "intuition" strategy is overwhelmingly popular and deeply unsatisfactory. And my own experience with high-powered solubility theory is that it is often a long-winded way of getting nowhere. However much of it I read, I could seldom use it to solve a problem.

In writing this book I am doing what I always try to do: finding the minimum amount of theory that gives the maximum amount of benefit. In the case of solubility I have found that five bits of theory give me this mini-max. Four of these bits of theory are well-known. The fifth is little known and yet will feature strongly in this book because it is of huge practical importance. Here is what lies at the heart of this book:

- Ideal Solubility theory for a crystalline solid simply tells you (mostly from its melting point) the maximum solubility you are likely to have without special effects. Simple and useful, and amazingly little-known.
- Hansen Solubility Parameters (HSP). Many readers will have heard of these. They are a 50-year old way of determining the best match between solutes (polymers, crystalline solids, nanoparticles, pigments ...) and solvents by providing three numbers for each solute or solvent which describe the van der Waals, polar and H-bonding capabilities. HSP are usable by experts and non-experts and cope well with the messy realities of a formulator's life. They have never gone out of fashion but recently they have experienced a boom in popularity, partly due to the HSPiP software of which I am co-author.
- COSMO-RS. The COnductor Screening MOdel - Realistic Solvents approach relies on deep thermodynamics and one-off quantum chemical calculations to allow highly accurate solubility (and related) predictions. One of the many reasons for the popularity of COSMO-RS is that even for those who do not fully understand how it works, the key components of COSMO-RS have a strong visual element that helps explain where the numbers are coming from.
- DLVO. This is the standard theory for dispersions and along with the idea of zeta potential and steric hindrance it is the only particle-specific tool that I have found to be of any practical benefit. There is universal agreement that DLVO has many faults and it is highly unlikely that (a) we know the exact inputs and (b) that the outputs are correct, but the general principles certainly help us to create some order out of a very complex system.
- KB theory. Kirkwood-Buff theory is the one that most people have never heard about. If you read the original 1951 paper it is highly likely that the 4 pages will mean as little to you as they did to me. However, it turns out that KB is rather easy to grasp and *much* more intuitive than all those things about

entropy, enthalpy, fugacity, standard states etc. which many of us were taught at some time and which most of us forgot as soon as possible. The real justification for including KB here is that from some rather trivial experimental data it is possible to resolve debates that have been raging for decades about solubility and solubilization, especially in the context of more complex systems in water where "water structure" and "hydrophobic forces" turn out to be as unnecessary as they have been confusing.

Even though the five ideas are not all that hard, I make the assumption that, like me, you cannot just see a formula and immediately grasp what it means. I have therefore linked the book to my large collection of "apps" that, wherever possible, do all the hard work for us so we can see (just by messing around with sliders) how the different parameters affect the outcome and, at the same time, provide the numerical answers you might need for a specific problem.

The apps run on all modern browsers, on phones, tablets and laptops, are free, free from ads and spyware and are safe to run on corporate networks because they are standard HTML5/CSS3/JavaScript. The code is all open source and I am always delighted to fix any bugs or add extra features if they will be helpful to someone.

If this were a Wikipedia article it would have lots of [citation needed] dotted through the text. When I write books, rather than academic papers, my choice is for a "reference lite" approach. References that seem to me to be key are provided and I add the occasional footnote if it seems appropriate. If anyone needs a citation for anything that I say, just let me know and I will be happy to provide it.

Some readers will be surprised to find that there is no chapter on crystallisation, which depends on solubility and yet there is a chapter on diffusion which, to many people, seems to be unrelated to solubility. I seriously considered crystallisation but in the end I have nothing useful to say as it is all about delicate balances of supersaturation, seeding, polymorphs etc. for which I can find nothing that can be put in a pragmatic app. Diffusion, on the other hand, makes little sense without understanding how it is strongly dependent on solubility issues and diffusion apps can teach the practical formulator about many things.

To my surprise, the final chapter is about the language we (mis)use to talk about solubility. The more literature that I read in order to write the book, the more I realised that many of the problems of solubility science arise from the way we frame the questions. By asking a question in a manner that pre-supposes one type of answer, many superior ways of getting an answer are automatically blocked. I believe that the approach here asks questions in a manner that is far more likely to produce actionable answers.

The downside to my approach in this book is that I might not cover some solubility area of interest to a specific reader. One of the many advantages of

publishing a free eBook is that I can respond quickly to feedback. Obviously I'm happy to fix errors and typos. But if you want a section or chapter on some solubility issue not covered here, I'm happy to see if it's the sort of thing I can learn adequately and for which a set of apps would bring the ideas to life. Any help is always gratefully acknowledged.

Acknowledgements

Which brings me to the acknowledgements.

My partners in HSPiP, Dr Charles Hansen and Dr Hiroshi Yamamoto have each taught me an immense amount about solubility issues. Our arguments have sometimes been heated not because any of us wanted to prove we were right but because we wanted to find out what was right. And all arguments had to stop when it was time for a bottle of champagne and dinner with some interesting mix of English, Danish and Japanese cuisine. This book would simply not have been possible if we had not embarked together on the HSPiP adventure and I thank them as friends and colleagues for all they have done for me.

My thanks to Prof Andreas Klamt and Prof Werner Kunz are straightforward. They are each outstanding pioneers in their fields. Andreas invented and commercialised COSMO-RS and is the CEO of COSMOlogic as well as Visiting Professor at U Regensburg. Werner is a Professor at U Regensburg and features in three major reviews quoted in this book: hydrotropes, Hofmeister and ionic liquids. I have benefited greatly from my interactions with them and I offer my sincere thanks for all the help they have provided. Any faults in my interpretation of their work are my responsibility.

My thanks to Dr Seishi Shimizu at U York have to be slightly more cautious as I do not want to get him into any trouble. Through his work on KB theory he has transformed the debate in many of the fields described here (proteins, Hofmeister, scCO2 entrainers, hydrotropes, cellulosics ...) and I have the honour of being a co-author on a few of his papers. So why do I have to be cautious? Seishi works in two modes: deep-theory mode to write unassailable statistical thermodynamical theory to be able to tackle complex problems; popular mode to take the results of those deep theories and apply them to questions that have perplexed formulators for decades. What I am doing here is writing in popular-popular mode, simplifying even further. This means that some of the ways I talk about some KB thermodynamics will not pass the Shimizu standards of precision.

So my thanks to him for his stream of new ideas, friendship, guidance and unbelievable patience (I've been a *very* slow student as I find this stuff *very* hard when I try to work at his level) is coupled with an apology (to him and to you as reader) if I don't quite get the wording right in some matter of statistical

thermodynamics. I acknowledge, therefore, my full responsibility for any KB errors in this text.

Happily, unlike the text, each of the KB-related apps *is* up to Shimizu standards. We have found that the best way to publish useful KB theory is to couple each paper with an app that brings the theory to life. Our way of working starts with me taking his formulae, implementing them in an app and then working together to fix my errors of implementation or, sometimes, to find ways to make the theory even clearer and more helpful. After a few cycles of refinement we end up with something that is both usable and correct.

And finally my thanks to many experts in the world of solubility who have invariably responded generously to my emailed questions and have helped me greatly to understand their points of view. My own efforts to give away my knowledge are a response to the many scientists who have been kind enough to help me learn from their insights and experience.

Steven Abbott
Ipswich, 2017

In April 2020, Lee McManus spotted many typos that have been in the book since its first edition. These have been fixed and I warmly thank Lee for the time and trouble needed to find them.

The Crystallization chapter was added in March 2021. It could not have been written without the vast knowledge of the crystallization literature generously provided by Prof Terry Threlfall. Anyone who reads the 19th century crystallization literature in German and then solves a major problem from that literature using 21st century techniques has my full admiration. I thank him warmly for his much-needed help. I'm also grateful to the team at SCM for providing access to the COSMO-RS capabilities of their Amsterdam Modelling Suite, which allowed me to show that even a non-expert such as myself could gather useful preliminary insights to guide a rational choice of solvents. At a key point while writing the chapter, Prof Werner Kunz provided some pointers to work and ideas I had completely missed, and I thank him, yet again, for his vast knowledge of many areas. Prof Alexander Van Driessche at U Grenoble Alpes has not only seen 12 x 4 m gypsum crystals first-hand in the Naisca mine in Mexico, he also pointed out some flaws in my thinking on clusters which I hope I've managed to fix. As always, responsibility for the remaining flaws is mine.

Abbreviations and Symbols

API	Active Pharmaceutical Ingredient
AUC	Analytical UltraCentrifuge or UltraCentrifugation
BCF	Burton, Cabrera and Frank crystal growth theory
CMC	Critical Micelle Concentration
CNT	Crystal Nucleation Theory
COSMO-RS	COnductor Screening MOdel - Realistic Solvation
cryo-TEM	Cryogenic Transmission Electron Microscopy
DLVO	Derjaguin and Landau, Verwey and Overbeek theory
DFT	Density Functional Theory
DOSY	Diffusion-Ordered NMR Spectroscopy
EVOH	Polyethylene vinyl alcohol
H-bond	Hydrogen bond
HSP	Hansen Solubility Parameter(s)
HSPiP	Hansen Solubility Parameters in Practice
HT	High Throughput (robotic) techniques
IL	Ionic Liquid
KB	Kirkwood-Buff theory
KBI	Kirkwood-Buff Integral
kT	Boltzmann Constant times Absolute Temperature – thermal energy
LP-TEM	Liquid Phase Transmission Electron Microscopy
MAC	Minimum Aggregation Concentration
MD	Molecular Dynamics
MHC	Minimum Hydrotrope Concentration
MPt	Melting Point
MSR	Molar Solubility Ratio
MVol	Molar Volume
MWt	Molecular Weight
NADES	Natural Deep Eutectic Solvent
NOESY	Nuclear Overhauser Effect Spectroscopy
NTA	Nanoparticle Tracking Analysis
PE	Polyethylene
PM(E)MA	PolyMethyl (or Ethyl) Methacrylate
PNIPAM	Poly N-isoPropyl Acetamide
RDF	Radial Distribution Function
RT	Gas Constant times Absolute Temperature – thermal energy

SAXS	Small Angle X-ray Scattering
SANS	Small Angle Neutron Scattering
scCO2	Supercritical Carbon Dioxide
SNT	Secondary Nucleation Threshold
TEM	Transmission Electron Microscopy
TMAO	Trimethylamine N-oxide (also called TMNO)

1 Solubility Basics

Solubility seems such an easy concept that it is hard to imagine that anyone would bother to write a book about it. You add a solute to a solvent, give it a stir and it dissolves. Or you add solute to a hot solvent, give it a stir, dissolve it, then let it cool and watch nice crystals fall out of solution. Or you have a nice solution and add another solvent to it and a solute falls out. How hard can it be to provide the scientific tools to make sense of solubility?

At university I was taught that it is so hard to make sense of solubility that I shouldn't waste my time trying to do so. And every time I came across the thermodynamics of solutions I was reminded why that was such good advice. If you open just about any book on solubility thermodynamics you will find that the formulae (of which there are many) are not especially hard. They are generally confusing because they are full of subscripts and superscripts, 0's and *'s that make subtle distinctions between one thing and another. But they are not in themselves hard. The key problem is that 10 pages later you are none the wiser. To me, thermodynamics is like a bad joke that never seems to reach the punchline. It's not so much that it is hard for most of us but that it seems to be pointless.

Worse than that, the slightest error in a subtle distinction and everything that follows is erroneous. You may *explicitly* know the distinction between constant pressure and constant volume, but it is easy to make an *implicit* assumption and end up totally wrong. Highly-trained thermodynamicists have come unstuck by, for example, confusing (perhaps implicitly) molarity with molality.

Faced with this, the vast majority of those who have to cope with solubility issues use vague terms like "hydrophilic" versus "hydrophobic", or "polar" versus "non-polar". To demonstrate how useless these terms can be, here is a (slightly modified) quote from an academic paper: "X is hydrophobic because it is insoluble in water so we therefore dissolved it in ethanol". If their definition of "hydrophobic" means something soluble in ethanol, what term do they use for something that is soluble in heptane?

Other attempts involve the sciencey $\log K_{OW}$ (logP) the octanol/water partition coefficient. This single number has near mythical status, yet it is absurd to think that you can encapsulate much of importance about a molecule via a single number. In another context I wrote the sentence: "To take an example at random, the $\log K_{OW}$ values (found at ChemSpider) of ethyl iodide and terephthalic acid are both 2.0, yet they are different in so many ways: aliphatic vs. aromatic, halogenated vs. non-halogenated, liquid vs. solid at room temperature, non-polar vs. polar".

There are plenty of alternatives such as Kauri-Butanol Number (don't ask), Kamlet-Taft (3 parameters that regularly appear in correlations but which have

never proved of decisive value), and we will later analyse some of the more popular schemes that never quite made it. The problem is that this babel of ideas creates its own confusion and most of us simply give up.

I started to get seriously interested in understanding solubility via a mix of desperation and chance. The desperation was due to some intractable (to me) solubility issues with some polymers being used in coatings. The chance was meeting Dr Charles Hansen who had developed Hansen Solubility Parameters, HSP, some decades earlier. HSP turned out to be amazingly useful in solving my immediate problems and I grew to use them over a wide range of applications. The theory behind them ("regular solution theory") is simple enough that I could get the general idea without having to fuss too much over the details, yet deep enough to have at least a reasonable grounding in real thermodynamics.

After a while I got reasonably comfortable with the few approaches that provide at least some utility to the solubility world: HSP, UNIFAC (and its variants), NRTL-SAC, Abraham Parameters, COSMO-RS. In practice, UNIFAC is too expensive for most of us as the vast array of required parameters are accessible only to those with deep pockets - though from time to time I have used the public domain parameters that work adequately for many molecules. NRTL-SAC and Abraham Parameters, for all their elegance, don't seem to have a broad-enough user base to have flourished. COSMO-RS will be discussed at length as it is immensely powerful and built on some concepts that are at the same time thermodynamically rigorous and chemically intuitive.

But then I stumbled across "solubilizers" – molecules that aren't especially good or useful as solvents but which can increase the solubility of solutes in other solvents – especially water. These seemed to lack any useful predictive tools and it quickly became clear that my favourite tools, HSP and COSMO-RS were useless in this context. Indeed, the whole area of solubilizers turned out to be confusing because everyone was using similar words (such as "hydrotrope" or "solvosurfactant") to describe these effects, whilst the word "hydrotrope" was also used to describe unrelated ideas (such as skin-friendly surfactants or microemulsions). Again we have solubility confused by a babel of languages.

That is when I came across Kirkwood-Buff (KB) theory. Like everything else in solution thermodynamics it is as dry as dust and seems to have no connection to reality. I would quite happily have remained in total ignorance of KB. However, the key idea of KB theory is conceptually simple and, unlike the rest of solution thermodynamics, can be used across the whole spectrum of solubility issues – from the mixing of liquids to the folding of proteins, from dilute to concentrated solutions. Even better, some rather simple (though admittedly tedious) measurements give you all the information you need to untangle what is going on. The theory is "assumption free", so is not built on a bunch of idealised concepts (such as infinitely dilute solutions) that tend to be the basis for much of the thermodynamics we are supposed to be using.

Despite its great power KB theory is almost unknown. One of my missions in this book is to popularise KB. So my rendition of KB will be bare bones, omitting many of the niceties that occupy the lives of professional thermodynamicists. The niceties *are* important, but they get in the way of the core message and my concern is the core message. I am grateful that there are brains far smarter than mine that enjoy the intellectual challenge of dealing with abstract thermodynamics. My job is to take their accomplishments and show how we can all use KB to produce clarity to solubility issues which have often been mired in decades of arcane thermodynamic disputes.

Let me give you two key examples.

First, the many oddities of water as a solvent have produced endless papers on why solute X will be more soluble in the presence of hydrotrope Y or why protein A is more or less folded[1] in the presence of additive B. The mysterious "water structure" has been invoked endlessly as an explanation, shedding, as it happens, no clarity at all on the matter. A few KB calculations, on the other hand, can reveal from rather simple experimental data exactly what is interacting with what and therefore causing the increase in solubility or the change in folding. So far, the results show overwhelmingly that "water structure" is of minor or even zero importance.

Second, there continues to be endless debates on entropy/enthalpy compensation. The heart of the debate is that most of us feel comfortable with enthalpic effects – we can understand why a molecule may be in a higher or lower energy state in the presence of a different molecule – but feel uncomfortable with entropic effects. Annoyingly, data often suggest that an intuitively-understood change in enthalpy is offset by a compensating effect in entropy; in other words, if there is stronger enthalpic binding, which is good for solubility there is a decrease in disorder, so entropy is decreased which is bad for solubility. The reason that this problem arises so often is that much of the thermodynamics of solubility (including HSP) is based on enthalpic arguments, yet the frequent apparent compensation by entropy renders enthalpic arguments useless. KB, however, works naturally in the free energy domain and describes molecular interactions in a manner such that entropy/enthalpy "explanations" do not arise. The problem with entropy and enthalpy is not that they are wrong, but that they are too crude; they are bulk values whereas we are interested in molecular explanations, which KB produce very naturally.

Having given you an outline of what I want to do, let us get going on the one bit of modestly hard work in the entire book by introducing the core concepts. Each concept is rather straightforward and an app is always at hand if needed to bring ideas alive. Remember that this stuff is so easy that even I understand it.

---

1 The world of protein/polymer effects uses numerous terms (e.g. coiled, extended, globular, native, denatured) in ways that I find confusing. I find the neutral terms "folded" and "unfolded" to be much more useful and have used these terms as much as possible.

1.1 The core concepts

1.1.1 Concentration

Thermodynamicists seem to love to torture us by shifting between concentration terms for no apparent good reason. They also shift between symbols for concentration. Although using just one concentration scale might seem attractive, it quickly becomes clear that two scales are necessary.

x is Mole Fraction. The great thing about mole fraction is that we know that it always goes from 0 to 1. When you have added no chemical its mole fraction is 0 and when you have the pure chemical its mole fraction is 1. This means that graphs with mole fraction along the X axis allow automatic comparison of what happens with any chemical.

c is Molar concentration, Molarity, in mol/l. This is the real world – how much stuff we actually have. When x=1 for water we have c=55.6 mole/l, because we have 1000g of water and its MWt is 18, so c=1000/18 mol/l. When x=1 for acetone we have c=791/58=13.6 mol/l. You can see why x is so clean – a graph of water concentration would go from 0 to 55.6 and that of acetone concentration would go from 0 to 13.6, but in terms of x both go from 0 to 1.

Of course we sometimes have to use Wt% for real world, but that will not appear in any thermodynamic equations. Surprisingly (to me) Volume %, φ is thermodynamically meaningful so appears from time to time in this book.

There is another solubility term that can appear in thermodynamics, Molality, which is solubility in g/1000g and is usually shown as m. Because the word is so similar to Molarity it was easy to imagine that it was invented purely to torture those who had to study thermodynamics. Actually it is rather a sensible unit. As we will see, the way that volumes change during solvation (another way of saying that densities change) are crucial to understanding the processes involved. This has the unfortunate side effect of leaving you unsure what c is at any point. You will always know how many moles you have at any given mix, but you cannot be certain of the volume (it makes no difference if you added known volumes or known weights) so the molarity is uncertain. If you work in molality you never have these uncertainties as you always know the weights of everything present. Despite this distinct advantage, few of us habitually use molality so m will not appear again in this book.

If you read thermodynamic papers you will often find n (or N) as a measure of concentration. This is the "number density", i.e. the number of moles per unit volume. In other words, n is a confusing way to talk about c. One equally finds ρ used for this "number density", which is even more confusing given that the actual density, ρ, is often required; ρ is used here exclusively as density. Because thermodynamicists regularly speak a confusing language I should

mention that the term "co-solvent" is often used to describe the addition of a solid such as urea or sucrose. Having got used to the term I can see why they might want to use it, though it is very unhelpful to the general reader.

1.1.2 MVol

The next term is really easy: Molar Volume, MVol. This is the volume occupied by one mole of the chemical. Calculating it is *apparently* straightforward from the molecular weight, MWt and density ρ:

$$MVol = \frac{MWt}{\rho} = \frac{g/mol}{g/cc} = cc/mol$$

Equ. 1-1

As I find it useful to check the “dimensions” of an equation, MWt is in g/mol, density is in g/cc so MVol is cc/mol. Of course I could use real units, kg/mol and kg/m³ and express MVol in m³/mol. One of the huge problems in thermodynamics is units. MWt is universally expressed in g/mol so that is why it is common to use density in g/cc and end up with MVol in cc/mol. Should I have said g/mole instead of g/mol? In the early draft I fluctuated between the two and decided that mol is cleaner.

The use of “apparently” was a warning about another thermodynamical nicety. The MVol of something that, out of solution, is a solid is *not* MWt/ρ_{solid}. Instead the density is that of a “virtual liquid”. This is annoying, but makes sense. When the solid is dissolved in the liquid its “density” excludes all the special interactions that make it a solid in the first place, hence we need a virtual density. Even if we don’t know the virtual density, at least we know that it will never be higher than its solid density, so we have *some* idea of what the MVol might be.

Who cares about MVol? Surprisingly MVol and ρ are hugely important for grasping what is happening within a solution. Although thermodynamicists love to talk about “partial molar volume” the word “partial” adds no value to our discussions, so we will just use MVol. The phrase is used in the context of mixtures of chemicals where the effective density of each chemical is affected by the local environment. So if we have chemicals 1 and 2, and chemical 1 is present at mole fraction x, then $MVol_1$ at concentration x and $MVol_2$ at concentration (1-x) are “partial molar volumes” to thermodynamicists and “MVols” to us. In each case they are MWt/ρ; it is the ρ of each component that changes across the solubility range.

My fight against unhelpful nomenclature will continue throughout this book. While writing it I had to read a draft of a paper on COSMO-RS theory. It was fascinating but I was thrown by the fact that it was concerned only with the

“residual” energy of the system. Why wasn’t it concerned with the “main” energy of the system? It turns out that what you and I would call the main energy, thermodynamicists decided to call “residual” because it was the residue when you removed some standard boring terms of no interest to anyone. You can see why a thermodynamicist might want to name the molar volume “partial” when the chemical of interest is only part of the whole, or "residual" when it is obvious to them that it is the significant residue, but it doesn’t add useful clarity and certainly adds confusion.

1.1.3 μ

The next concept is μ, the *chemical potential,* in units of J/mol*.* Those who have met μ before will be aware that almost immediately it gets festooned with superscripts and subscripts: μ^*_o and such like. There are very good reasons (the problem of “standard states”) why the careful thermodynamicist must festoon μ with such things, but for our purposes we just need plain μ. We are familiar with potential energy in physical objects. A ball at the top of a hill has a positive potential energy and will happily roll down the hill. A ball at the bottom of a valley has a negative potential energy and the only way it can move is if there is an even deeper valley or if some other form of energy is applied to lift it up. Chemical potential is exactly the same. A system with a high, positive, chemical potential will want to lose that energy and change its state. A system with a large negative chemical potential is highly stable.

An internet search shows plenty of articles starting “Most chemists do not understand chemical potential so here we explain how simple and useful it is” continuing with long explanations that convince me that it is neither simple nor genuinely usable by most of us. I cannot avoid using it, but as long as you get the general idea, that is all you need. The relationship of μ to more familiar concepts is described in the following two sections.

1.1.4 G

Free energy, G, is similar in concept to chemical potential though it represents the whole system rather than just the “chemical” part of the system. We have all been taught that large negative G is stable, and large, positive G is unstable. This simple statement happens to be false, so we shall come back to this in a moment and see why chemical potential is so much more useful. It is deeply unfortunate that the key variables, to be described later, in KB theory are also called G_{ij}, G with a pair of subscripts; these are the KB Integrals, KBIs. I will try to enforce the rule that all plain G values are free energy and all double-subscripted G_{ij} values are KBIs.

When I spoke about hills and valleys I omitted to mention the flat, neutral, planes. They are a sort of reference level and if this were a real thermodynamics book we would now spend a lot of time discussing “standard states”. If I live

on the Tibetan plateau, I might regard that as my standard state because it is easy to walk downhill from that plateau and hard to walk up the hills above the plateau. Yet to those of us who live near sea level, the Tibetan plateau is very far from our standard state as it requires a large amount of energy to reach it. If you are doing precise thermodynamics it is vital that everyone agrees on the standard state. In this book we will just assume that the standard state is "obvious" and discuss it no more. This is not entirely disreputable. My version of KB works because high-powered thermodynamicists have done all the hard work to get the definitions correct. I greatly appreciate their hard work so that you don't have to expend the effort.

Now it is time to see how G and μ are inter-related:

$$\mu = \frac{\partial G}{\partial c}$$

Equ. 1-2

where c is the concentration.

We shall see that when it comes to mixing two chemicals, this equation is super-important. It is often said that a mixture is stable if its free energy is negative. Because of the problem of standard states, this statement is plain wrong. The definition of stability is that $\partial G/\partial c$ is negative (i.e. a negative chemical potential, as discussed above) and that $\partial^2 G/\partial c^2$ is positive (a more subtle point I won't discuss further).

The fact that μ and G are related via a derivative is deeply unfortunate. Knowing a single free energy tells you nothing about the chemical potential. Although KB theory is assumption-free, we invariably end up having to use derivatives of things like free energy, which means that we need multiple points in concentration space from which the derivative can be derived either as $(G_{@c1}-G_{@c2})/(c_1-c_2)$ or by fitting the data to (say) a polynomial and calculating the derivative. If the polynomial is $x+y.c+z.c^2$ then the derivative at c is $y+2z.c$.

1.1.5 a and γ

We cannot avoid a and γ, activity and activity coefficient. Activity is the effective concentration (or whatever is the key measure of interest) of a solute and activity coefficient is the ratio of activity to real concentration. So if the real mole fraction concentration is 0.15 and the solute behaves (e.g. by having a higher vapour pressure) as if it has a concentration of 0.3, a=0.3 and γ=2.

$$a = x\gamma$$

Equ. 1-3

Because so many of us are uncomfortable with chemical potential, it is a relief to know that it relates directly to activity, with which most of us are comfortable:

Equ. 1-4 $$\mu = RT\ln(a)$$

Another pair of terms which often appears are “fugacity” and “fugacity coefficient”. Real thermodynamicists are happy to use them in places where I use activity and activity coefficient, but as the terms are much more associated with vapours “fleeing” from a liquid and because in this book we are generally interested in solubility of crystalline solids, polymers, nanoparticles etc. I won’t use the term again. If, in general reading, I substitute “activity” for “fugacity” I generally find that it does little harm.

1.1.6 And a few other things

We have already mentioned the controversy around enthalpy and entropy. By using KB we can escape the largely fruitless debate and will not have to do any calculations featuring the two properties. However, when it comes to day-to-day useful techniques such as Flory-Huggins theory of polymer solubility we will note in passing that some of the terms cover the necessary entropic parts of the theory. Fortunately the use of entropy is so simple (perhaps simplistic) that we can describe the terms very easily and use the formulae with confidence without getting mired in any of the subtleties.

Because MVol and density are both necessary, we will see that pressure, P, can be a useful parameter, in particular with scCO2 (supercritical carbon dioxide). A lot of thermodynamics involves P but because most of us work at 1 bar, and because significant P-effects require 100 bar pressures, we miss out on a lot of good stuff except when we *have* to use high pressures for scCO2. Because we do not have to involve P outside scCO2, we will assume that our intuitions about P are reliable and it will be discussed no more.

There is another type of pressure, Π, the osmotic pressure. We can regard this as the physical manifestation (i.e. you can measure it with a pressure gauge) of the chemical potential. So by measuring Π in some systems we can find out about μ and via KB can grasp what is going on with our solute and solvent molecules. As most of us have not seen an osmometer since (perhaps) college days, Π will appear only rarely in what follows.

1.2 KB theory

Normally before embarking on the theory of statistical thermodynamics, the reader would be urged to be patient because despite the pain or tedium it will all make sense. With KB there is no need for pain or tedium[2]. It is all rather

2 I have endured much pain and tedium in trying to understand KB sufficiently well to provide a painless explanation. Thermodynamicists have minds that thrive on abstract entities and cannot comprehend why people like me do not find their abstractions as meaningful as they do.

straightforward and the key concepts link exactly to the intuitions of anyone who has played with solvents and solutes.

Although KB are painless, I still need to explain why your life will be improved if you go to the trouble of reading through what follows. So I offer you a guarantee. I guarantee that armed with the intuitive ideas behind KB and numbers called Kirkwood-Buff Integrals (KBIs), you will be able to understand a huge variety of solubility issues with clarity and ease. As we shall find throughout this book, many solubility puzzles are solved via simple arithmetic: "this KBI is large, this one is small, this one is irrelevant so, therefore, the solubility effect is explained by the large KBI." It doesn't get simpler, or more powerful, than that.

Let us start with a simple observation. A molecule doesn't know if it is a solvent, a solute, a hydrotrope or anything. All it knows is that it is surrounded by other molecules. So although we will give molecules labels such as A and B or 1 and 2, they are only for our convenience and there are no special rules (such as the need for infinite dilution) to restrict what we are doing.

Now imagine yourself looking out from one of the molecules at the molecules around you. They are constantly jiggling around due to thermal motion and it is a bit confusing at first, but over time it is clear that there are three very simple rules. The first two rules are the same whether you are surrounded just by molecules identical to yours (i.e. this is a pure solvent) or have one, two or many different types of molecule around you. Here are the three rules:

1. Any molecule (including the central one) in a given position means that there is an exclusion zone extending over the volume of that molecule plus the volume of other molecules – i.e. molecules cannot overlap. This is trivially obvious. And yet already you can see that MVol is going to play a key role in how molecules can interact. The "excluded volume" effect is often the simple explanation for effects that have caused decades of confusion, especially in the case of proteins and other polymers.
2. It is possible to count the average numbers of different types of molecules in the immediate neighbourhood then stretching out far into the bulk of the system. Note how general this statement is. It is not at all the same as saying "We can look out and see a solvent shell". If you have not met the phrase "solvent shell" then life is easy for you. If you have met it then the chances are that you have some clear idea of what it means. This (and I speak from my own experience) idea is almost certainly false. Just forget about "solvent shells" and think about this extended sea of molecules, each of which you can count, from nearby into the far distance.
3. You will notice that on average some types of molecules, for whatever reason, tend to prefer to be together. We are not talking about specific interactions such as complexes, simply of a general trend for there to be more, say, 1's near 2 than would be expected by chance.

From those three rules the whole of KB theory follows. Because if we can know how much 1 prefers to be with 1 than with 2, or how much 2 likes to be near 1 if 3 is also present then we can understand why 1 is a good (or bad) solvent for 2 and whether 3 can help or hinder the solubilization of 2 in 1.

In principle we can *calculate* the results of the three rules. In practice this is currently impossible, even for those with a spare supercomputer. The reason is not that the calculations are especially hard. Rather it is because the outcome is the balance of many small interactions and the slightest error in the estimation of the interactions (e.g. a slight mis-choice in the force field of a molecular dynamics (MD) calculation) is enough to tip the balance in the wrong direction.

Happily we can *measure* these things via rather simple (though generally boring) techniques. So although we cannot achieve *prediction* we can achieve *understanding* which is almost as good. The only reason I am spending any time with KB is because it is a path to enlightenment absent from all the other thermodynamic theories that are out there. KB theory was developed in 1951[3] (the paper is only 4 pages!) but was not much used because the key values could neither be calculated nor measured. It was Arieh Ben-Naim who first worked out how to derive the values from (simple) experimental measurements and those who want to dig deeply into KB should read Ben-Naim's wonderful book *Molecular Theory of Solutions*[4].

At last we can reveal a typical KB formula for a mix of two molecules 1 and 2. The equation is taken straight out of the original KB paper:

$$\frac{\delta\mu_1}{\delta c_1} = RT\left[\frac{1}{c_1\left(1+c_1\left(G_{11}-G_{12}\right)\right)}\right]$$

Equ. 1-5

This tells us that the change (per mole) of chemical potential of molecule 1 with respect to the (molar) concentration of 1, c_1, depends on the concentration of 1, on RT, the gas constant times absolute temperature and, finally on the so-called KB integrals G_{11}, and G_{12}.

What does this mean? The change of chemical potential in a system where the molecules aren't bothered whether they are associated with themselves or other molecules is simply RT/c. This is because in such a system all the G_{ij} values are the same so the terms cancel out. In systems where 1 prefers to be with 1 than with 2 then $G_{11} > G_{12}$ so the G_{ij} in the equation start to matter. The equation for the change of chemical potential with molarity is very natural for those of us who think in terms of molar concentrations. It is less natural for creating graphs

---

3 J. G. Kirkwood and F. P. Buff, *The statistical mechanical theory of solutions. I,* J. Chem. Phys. 19, 774-777, 1951

4 Arieh Ben-Naim, *Molecular Theory of Solutions*, OUP, 2006

in apps because there is no simple way to plot the whole solubility range from 0-100% of 1 (and 100-0% of 2). For plotting convenience, therefore, we use mole fraction. The equivalent equation (which contains a c_2 term but we can still do the plot across the x_1 range) is derived from the previous equation by a bit of algebra that need not concern us and provides us with a very nice way of thinking about the relative G_{ij} values:

$$\frac{\delta\mu_1}{\delta x_1} = RT\left[\frac{1}{x_1\left(1+x_1c_2\left(G_{11}+G_{22}-2G_{12}\right)\right)}\right]$$

Equ. 1-6

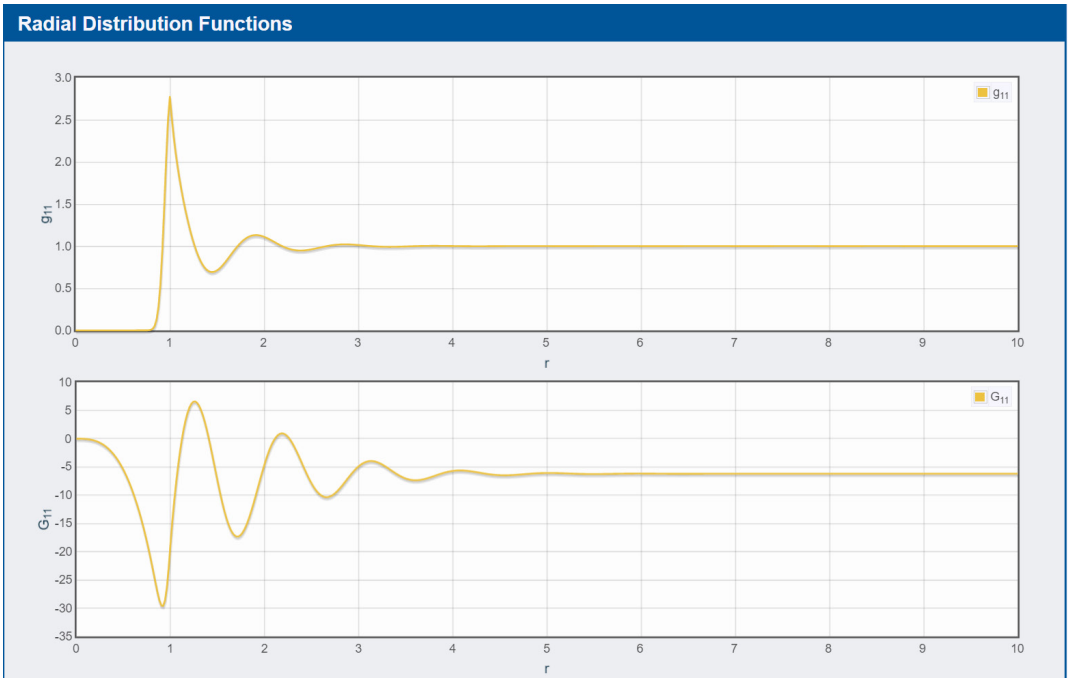
Now we have the important term $G_{11}+G_{22}-2G_{12}$ which sums up much of what you need to know about KB – it represents both of the self-association terms and how they compare to the cross-association term which is multiplied by 2 because it contains G_{12} and G_{21} which are identical.

What are these G_{ij} values? They are called KB integrals (KBI) and are formally defined based on the integral of the radial distribution function between i and j where i and j might be the same or different molecules. The radial distribution function (invariably shown as g_{ij}) can be considered as the relative number of molecules of type j around molecules of type i at distance r, compared to the bulk number. This immediately brings to mind ideas of solvent shells that some of us have been brought up on. Please forget these - they are deceptive and confusing. The radial distribution functions are statistical averages and are far better thought of in terms of relative densities, ρ_{ij}. This gives us the definition at radius r with respect to the bulk density:

$$g_{ij} = \frac{\rho_{ij}(r)}{\rho_{ij}^{Bulk}}$$

Equ. 1-7

It is helpful to get a feel for both g_{ij} and G_{ij} from a simple radial distribution function (RDF) app:



App 1-1 https://www.stevenabbott.co.uk/practical-solubility/rdf.php

From the definition based on densities (rather than our normal intuition of radial distribution functions counting j's around i's and i's around j's) it is clear that $g_{ij}=g_{ji}$. The *integrals,* G_{ij} are derived from the rdfs via the integral term, $G_{ij} = \int 4\pi r^2(g_{ij}-1)$, which simply describe the number of i-j pairs more (or less) than the average, using the -1 to remove the background level. It naturally follows that $G_{ij} = G_{ji}$ because the counts are pairwise. Again the *intuition* suggests that the number of i molecules around j's will be different from j molecules around i's, but the G_{ij} values are global averages throughout the solution, hence the equality of ij and ji. Note that the G_{ij} integration also involves the radius, r, via the $4\pi r^2$ term, because there is a bigger volume in which to count the numbers of molecules. You will notice that the relatively gentle fluctuation in the values of g_{ij} at higher radii are amplified in G_{ij} because of the $4\pi r^2$ term.

Fortunately it is very easy to grasp, in general, what these KB integrals mean. If, when there are similar numbers of 1 and 2 molecules, G_{11} is greater than G_{12} it means that 1 prefers to be near 1 rather than near 2. Indeed, G_{12} can frequently be negative, i.e. there are fewer 1-2 interactions than would be expected on average. So in our first equation a *larger* G_{11} will result in a *smaller* change in chemical potential which is the same as saying that 1 is less happy to go into a mixture with 2 than if it had no preference. When G_{12} is negative, which is often the case if G_{11} and G_{22} are positive then, because it appears as - G_{12} in the equation, that reinforces the fact that this is an unhappy mixture. If G_{12} is more positive then 1 positively likes being surrounded by 2's, so is happier to be in solution.

The same thing can be seen with the bigger picture that opens up when mole fractions are used. Now we compare $G_{11}+G_{22}$ with $2G_{12}$, in other words, are the

total attractions of 1 for 1 and 2 for 2 greater than the mutual attractions of 1 and 2?

Those who happen to be familiar with regular solution theory (or who come back here after we've discussed it in a later chapter) will see a parallel with the 11, 22 and 12 terms that are commonly discussed in that theory. The huge advantage of KB is that there are no approximations (there are many in regular solution theory) and no tying oneself in knots about "geometric means" and such like. But if you are already comfortable with the thinking behind regular solutions then KB presents no new problems.

There is one more number that can be calculated from the KB integrals. Some find it helpful and focus on it, others are less keen on it. Either way it is important to know about it. The number is the *excess number* N_{ij} which is the number of i's around j in excess of what would be expected from a random mixture, remembering as always that i and j can be the same molecule, so we can have the excess number of i's around i and j's around j via N_{ii} and N_{jj}.

The definition is:

Equ. 1-8

$$N_{ij} = c_i \left(G_{ij} - G_{ij}^{Ideal} \right)$$

Unlike G_{ij} the definition of N_{ij} (which has the c_i term) means that it is different from N_{ji} , so the excess number of j's around i is not necessarily the excess number of i's around j.

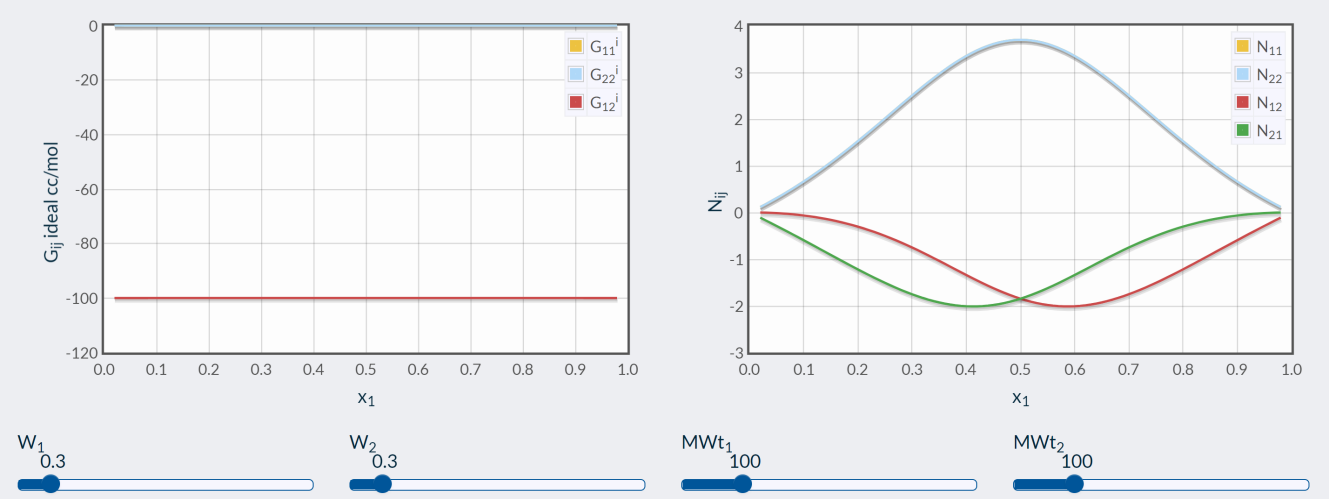
The definition also shows the problem with N_{ij}. Because it is an *excess* number it has to be referenced against a standard so it requires a definition of the *ideal* KB integral, i.e. the one where there are no special interactions. There is a large literature debating quite how to calculated that ideal value. The niceties do not concern us at present and as it is a number we can readily relate to we will use it from time to time and it features in the app.

It is very important *not* to think of "solvent shells" or "complexes". If N_{ij} =2 that does *not* mean that there are 2 i's sitting neatly next to a j and if N_{ij}=1 that certainly does *not* mean that there is an i-j binary complex. These are statistical numbers saying that on average each j is seeing N_{ij} more i's spreading out into the distance than it would have seen on average. This will make more sense when we explore these numbers in the app.

So far we have been building up intuitions about what these KB integrals mean. Now it is time to see them in action within the basic KB app. At this stage we shall ignore the complexities of densities (which form the top portion of the app) and look at the basics of KBIs.

1.2.1 Getting a feeling for G_{ij} and N_{ij} values

This app takes some simple, familiar inputs and allows the user to see all the relevant G_{ij} values. How they are calculated is discussed in the next section. For now, focus on the inputs and outputs.



App 1-2 https://www.stevenabbott.co.uk/practical-solubility/kbgs.php

The inputs are the familiar MWt_1 and MWt_2 and two "W" parameters. For those who happen to know about Wilson Λ_{ij} ("large lambda") parameters the numbers will make sense. For now treat them as ways to adjust the activity coefficients of the two solvents. When both W values are 1 then the solution is ideal - neither solvent is bothered by the presence of the other and the activity coefficients are 1 across the whole mole fraction range of x_1 from 0 to 1. Because the graphs are auto-scaling you might think that when the W values are both near 1 that there are massive changes in activity coefficient, but when you look at the scale you will see that they are hovering very close to 1.

If both W values are less than 1 then this is the common example of a solvent and solute that don't like each other and the activity coefficients are higher than 1. If (and this is rare, for example with chloroform/acetone which attract each other) both values are greater than 1 then the activity coefficients are less than 1. You can generate a huge variety of activity coefficient shapes just by playing with the W values.

The MWt parameters have no affect on the activity curves as these depend only on mole fraction.

Now look at the G_{ij} curves. Assuming you have the default values with W~0.3 and MWt~100 then the activity coefficients are large, i.e. 1 doesn't like being in 2 and 2 doesn't like being in 1. What does this look like in terms of KB integrals? Because there is nothing mysterious about them we can answer the question before looking at the graphs. 1's will prefer to be with 1's and 2's with 2's, so G_{11} and G_{22} will both be positive over most of the range, approaching 0 in pure 1 or pure 2. Because 1's dislike 2's, there will be less 1-2 interaction than expected from random mixing, so G_{12} (which, of course, is the same as G_{21} which we don't have to plot) goes negative and is most negative when there are equal numbers of 1's and 2's.

Keeping the same settings, then we find that the *ideal* G_{ij} values are uninteresting and finally we can look at the excess numbers N_{ij}. Here we find that N_{11} and N_{22} are identical (only one is visible) with about 1.5 extra molecules around at the maximum point of a 50:50 mix. N_{12} and N_{21} are different but symmetrical with about 1 fewer molecule of 2 around 1 or of 1 around 2. In all the apps you can use the cursor to get detailed readout from the graphs.

You can develop your intuitions by changing the input parameters. Altering MWt values changes the symmetry of the curves in ways that are not trivial to work out. Try, for example, doubling MWt_1. This decreases the maximum value of G_{11} because there is so much 1 in the mixture because of its size that the self-association through dislike of 2 does not make a huge difference. At the same time it increases G_{22} because 2s sees a lot more 1 in the solution so are on average closer together than they would have been.

If you change both W values to 2, giving strong solvent-solute interactions, then the curves all flip in a manner that should by now make sense.

We are almost there. There is one question avoided up to now. Although we can make some sense of whether N_{ij} is 0.1, 1 or 10 because we can imagine 0.1, 1 or 10 extra molecules, what are we to make of G_{ij} values? First, what are the units? They are cc/mol, the same as MVol (which will be discussed again shortly), but that still doesn't tell us what they mean. I look at them as describing how much of the volume of the solution is taken up by non-average local concentrations. When we get (as we must, but only briefly) to discussing Fluctuation Theory this view should make more sense.

It is worth taking stock before going further. Why are we bothering about KB integrals and excess numbers? Because, as my guarantee stated, they are a uniquely powerful way to look inside a solution to see what is going on. Although in these examples we already know from the activity coefficients that solvent and solute like or dislike each other, the activity coefficients give no numerical insight at the molecular level. This is common feature with most classical thermodynamics and one which causes a lot of problems. With KB we get intuitively useful numbers that can be compared and contrasted

between different systems and which map onto our chemical understanding of why two molecules may or may not like each other. We shall also see clearly that the effect of molecular size, which is otherwise hard to grasp, is rather straightforward because of "excluded volume" effects. This is already a powerful case for KB but is not enough to convince me to write a book based on them. We will find that when we have three components the insights from KB are not just nice-to-have, they are must-have.

1.2.2 Measuring G_{ij} values

As mentioned earlier, there are currently no good methods for reliably calculating these G_{ij} values from molecular details of real world systems, though modern MD systems are getting better each year. Because the ideas behind these values are rather easy to visualise, it is not so hard to get a general idea of the relative sizes of these terms – you will know in general if 1 and 2 are similar, dissimilar or even complementary (i.e. they have something like complementary hydrogen bond donors and acceptors). Such intuitions can work reasonably when there is approximately equal 1 and 2. But how does G_{11} compare to G_{12} when there aren't so many 1 molecules around? That is harder to intuit and of great importance because we are rarely interested in 50:50 solubility. We really do need to know the numbers across a significant (ideally the whole) mole fraction range so it is unfortunate that we cannot yet calculate them.

If that were the end of the story then this whole KB section would be a waste of time for those interested in real world solubility. What saves the situation is that we can, with modest effort, *measure* the G_{ij} values.

To understand how this can be so, we need to dig a bit deeper into what is going on. In the preceding app we *calculated* the KB integrals from some assumed activity coefficients. Because we know that these can be readily measured (e.g. via vapour pressure measurements) in the sort of simple two-solvent system implied, it is already clear that a chain of logic exists to go from activity coefficients to KB integrals. By deliberately ignoring the density inputs of the app we missed out another key requirement for calculating the integrals.

In the case of a two-component system we know that simply from looking at the rate of change of chemical potential, which in the app was derived from activity coefficients, we have all the information needed to calculate the *difference* between G_{11} and G_{12}. This is because of the definition of KB integrals seen previously, along with the fact that μ=RTln(a):

$$\frac{\delta\mu_1}{\delta c_1} = RT\left[\frac{1}{c_1\left(1+c_1\left(G_{11}-G_{12}\right)\right)}\right]$$

Equ. 1-9

To know the *individual* G_{ij} values we need a source of data other than activity coefficients, i.e. we need two equations for two unknowns[5]. This is the reason we introduced the idea of MVol. Merely by measuring how the MVol of one of the components changes with its concentration we can gain new information. How is this possible? If each component self-associates exactly as it does as a pure liquid then the MVol of each component will be unchanged. However, if, for example, there are strong 1-2 associations then the MVol of each component is going to be different. Because MVol is simply MWt/density, by measuring the density of the solution containing known molar amounts of 1 and 2 we can extract the MVol data. It should be no surprise that there is a connection between MVol and KB integrals - their units of cc/mol are identical.

Here is the formal set of relationships that link activity coefficients and MVols to the KB integrals. Experts will note the omission of the isothermal compressibility term which is ignored in this book because the term is insignificant in the context of the large effects that interest us in practice. Note, for reference, that these are the famous Ben-Naim KB inversion equations which made it possible to link KB to experiments:

$$G_{ii} = -\frac{1}{c_i} + \frac{c_j MVol_j^2 c_{Tot}}{c_i D}$$

Equ. 1-10

$$G_{ij} = -\frac{MVol_i MVol_j c_{Tot}}{D}$$

Equ. 1-11

$$D = \frac{x_i}{RT}\frac{\delta \mu_i}{\delta x_i} = x_i \frac{\delta \ln(a_i)}{\delta x_i} = 1 + x_i \frac{\partial \ln(\gamma_i)}{\partial x_i}$$

Equ. 1-12

This tells us that G_{11} and G_{22} are calculated knowing:

- the individual molar concentrations along with c_{Tot} which is the sum of the two concentrations
- the MVol of the other molecule at that specific concentration
- the parameter D which is related to the chemical potential or, equivalently, the activity or activity coefficient.

The mixed value G_{12} needs the total concentration and the individual MVols at the concentration of interest. D brings in the other experimental data, the activity coefficients. It is shown in its three equivalent forms as it is easy to get confused when reading the literature. Note that D does not have a subscript because it is the same value if calculated via 1 or 2. Note, too, as I have found to my

---

5 You might think that we have two activity coefficients to give us two datapoints. However, thanks to the Gibbs-Duhem relationship, if you know one coefficient then you know the other, so we only get one datapoint.

cost, that small errors in calculating D can cause the G_{ii} values to explode at small values of c_i because D values become the difference between two large numbers.

So now "all" we need is the $MVol_i$ values. Where do these come from? From density data, of course.

As it happens, almost no one measures densities of solutions. Why would anyone bother – it seems such a low-tech and boring property? It is entirely understandable but deeply unfortunate that most of us have no density data on any of the solutions of interest to us. If they were commonly available then extensive KB data mining could take place. One reason for writing this book is to encourage the solubility community to get into the habit of obtaining these measurements. Those who imagine (as I did) that density measurements require the weighing of an accurately measured 100ml of liquid will be pleased to know that fully automatic density measurements need a few µl of sample to produce values significant to 4 or 5 decimal places, more than good enough to extract MVol values and therefore to calculate G_{ij} values. Relatively low-cost in-line density meters are especially useful for high throughput (HT) KB determination.

Unfortunately we need three tedious steps to be able to extract the MVol values from density.

Knowing the density of the solution of 2 in 1 $\rho_{2\text{-in-}1}$ and the density of the pure solvent 1, ρ_1 we can calculate an *apparent* $MVol_2$ which is the naive value assuming that 1 and 2 do not affect each other:

$$MVol_2^{App} = \frac{1}{c_2}\frac{\rho_1 - \rho_{2in1}}{\rho_1} + \frac{MWt_2}{\rho_1}$$

Equ. 1-13

Then from the apparent MVol the real one can be calculated:

$$MVol_2 = MVol_2^{App} + \left(\frac{1000 - c_2 MVol_2^{App}}{1000 + c_2{}^2 \dfrac{\delta MVol_2^{App}}{\delta c_2}}\right) n_2 \frac{\delta MVol_2^{App}}{\delta c_2}$$

Equ. 1-14

Finally, knowing the molar volume of pure 1, $MVol_{1pure}$ the MVol of 1 can be calculated:

$$MVol_1 = \left(\frac{1000 MVol_{1\,pure}}{1000 + c_2^{\,2} \dfrac{\delta MVol_2^{App}}{\delta c_2}} \right)$$

Equ. 1-15

Admittedly all this is rather tedious. And I can assure you that doing it in real life without decent apps is genuinely tedious. It is such a long chain of reasoning with lots of chances for a minor error in one of the links in the chain which will then invalidate the results. Tedious, yes, but not especially hard. It is, after all, mostly high school arithmetic and attention to detail. Once the process is set up properly to go from experimental data to KB integrals and excess numbers, it is not hard at all.

And the gains are huge. By combining the data from density with the data from (say) vapour pressures, you have access to deeply meaningful, assumption-free statistical thermodynamic numbers which have the bonus of possessing an intuitive meaning that we non-thermodynamicists can relate to.

So now we can go back to the app and see what the density curves are telling us.

1.2.3 Information from density

Density seems such a humble measure that it is hard to imagine that a curve of density versus x_1 would reveal anything of interest.

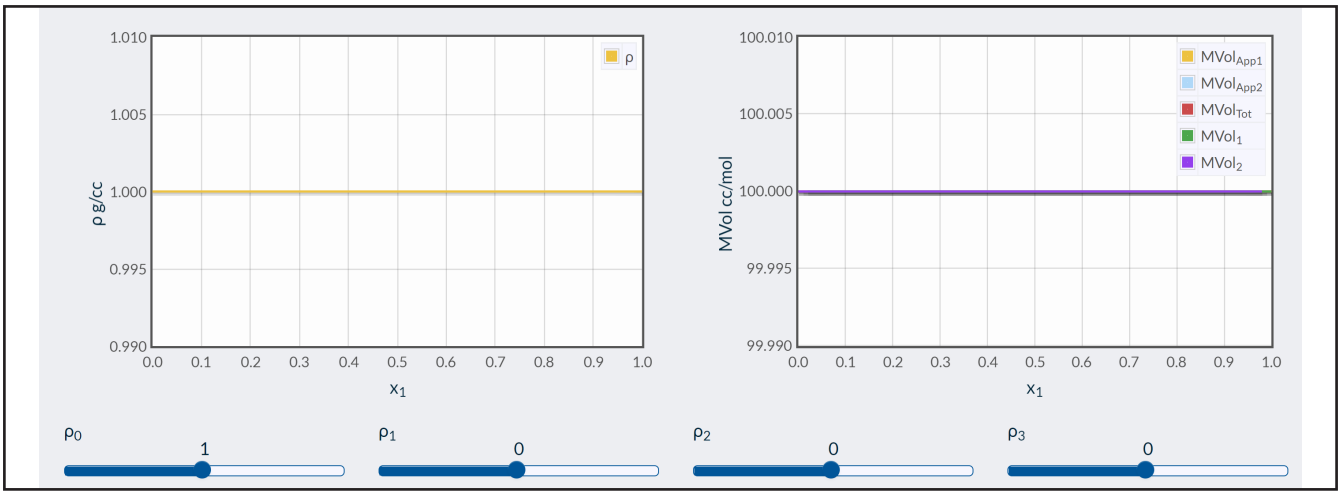


Figure 1-1 Nothing much happens with the density defaults

With the default, to make it simple to grasp the basics of the KB integrals, the density is a constant 1g/cc. The 4 ρ settings are for the polynomial $\rho = \rho_0 + \rho_1 x_1 + \rho_2 x_1^2 + \rho_3 x_1^3$. If 1 has a density of 0.9 and 2 has a density of 1.1 then ρ_0 is set to 0.9 and ρ_1 must be 0.2 - try it in the app. This gives a simple, ideal density

behaviour which by definition means that there are no interactions between 1 and 2. More interesting is when there is a complex density behaviour:

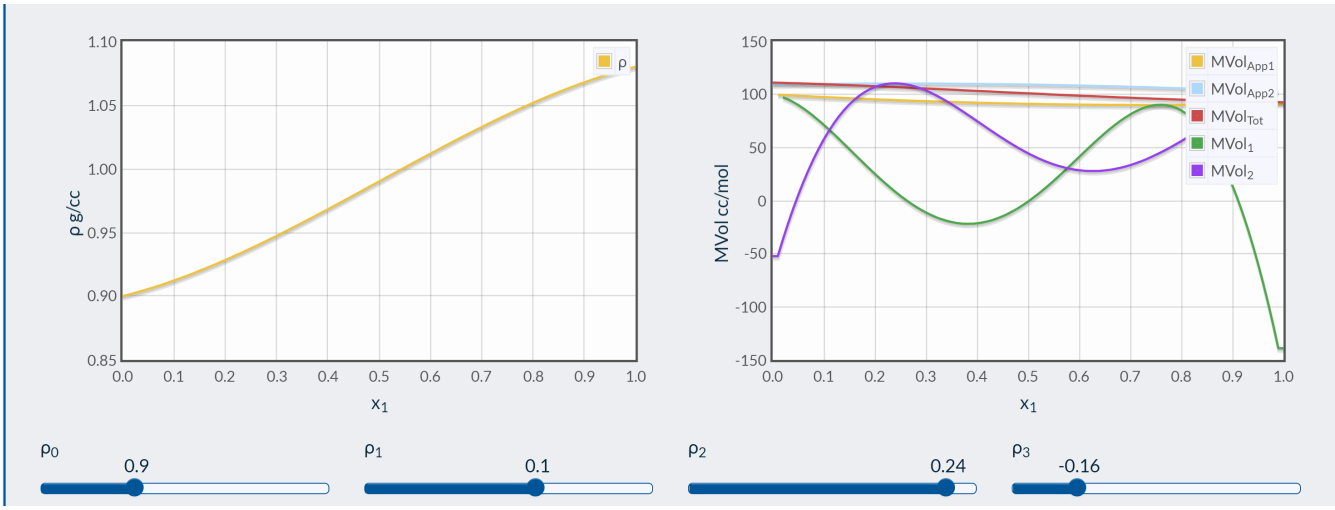


Figure 1-2 A non-linear density curve. By definition 1 and 2 must be interacting significantly

Here there are significant interactions because the density is not varying linearly from 0.9 to 1.1. If you just calculated the total molar volume M_{Tot} or the apparent values $MVol_{App}$ you would not think that much was happening - after all, the density curve isn't all that exciting - which is why no one bothers to measure anything so boring. Yet the real $MVol_i$ values are anything but boring. Look at $MVol_2$ at low values of x_1. It is a negative value - the molecule is taking up negative volume! This sounds wrong. But look at the initial slope of the density curve; as you add more high density 1 to 2 the density hardly increases. This can only happen if 2 takes up less space than it otherwise would have done.

1.2.4 Pressure

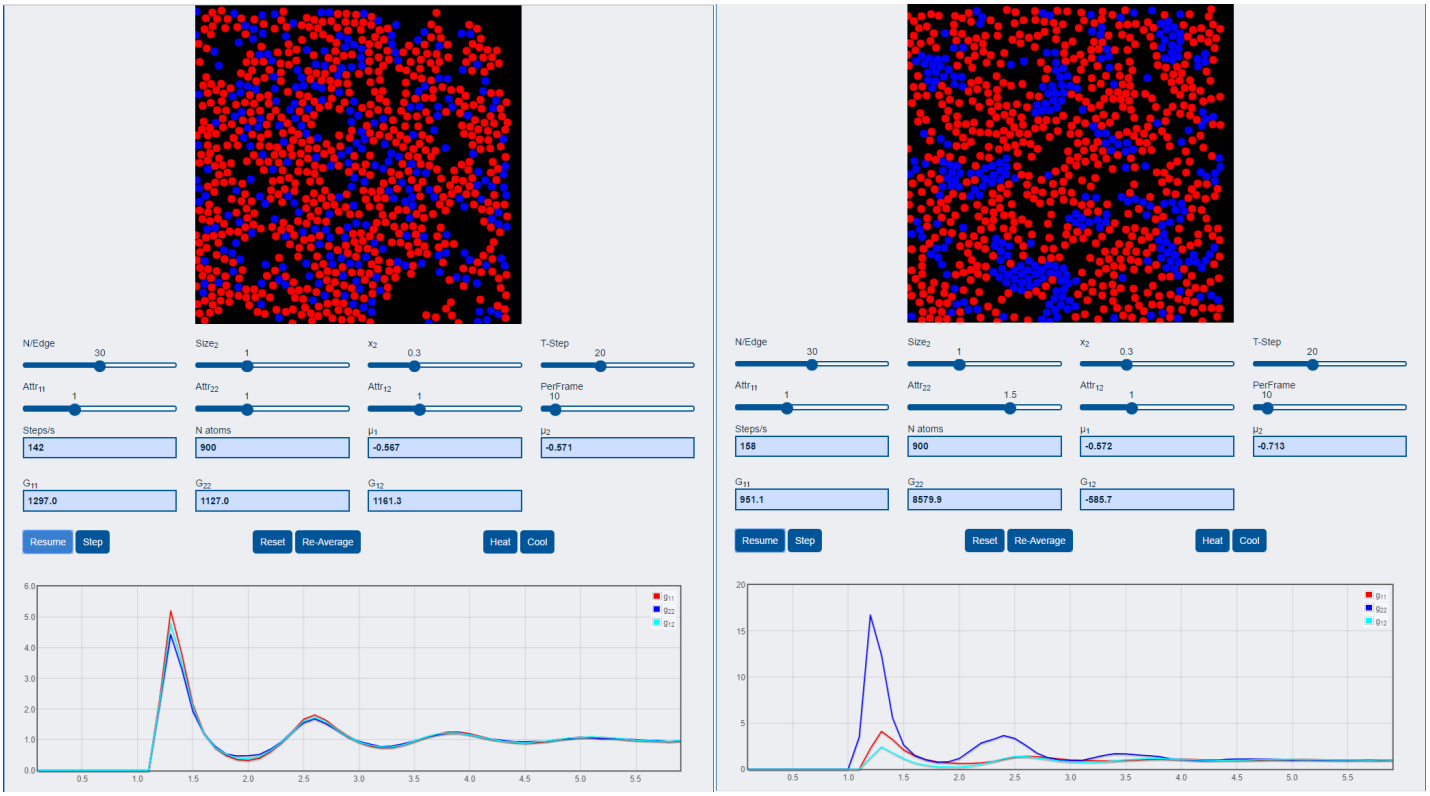
Pressure, volume and density are inter-related. So it is no surprise that G_{ij} values come from routine pressure-related measurements, i.e. from osmotic pressure Π, or by measuring the vapour pressure, each of which provides activity coefficients. Because the calculations always rely on derivatives (for MVol it is the change of $MVol^{App}$ with x and for D it is the change of γ with x) single measurements are of no use. In the old days measurement of densities, osmotic pressures or vapour pressures was a tedious process. With modern equipment there is no excuse for not doing measurements across the entire relevant mole fraction range.

Another possibility is to measure volume as a function of applied pressure and mole fraction. But earlier we simplified things by saying that we could ignore isothermal compressibility effects. This is another way of saying that to measure KB integrals via direct pressure measurements we need very high pressures that are not common in most labs. One exception to this is solubility in supercritical CO2, where pressures of a 100+ bar are routine.

This section is for intellectual completeness. Very few of us will play with pressure in our solubility experiments[6]. The day that someone comes up with a simple high throughput pressure system where we can measure vapour pressures and densities, then a whole new way of gathering KB data will open up.

1.2.5 A toy KB world

We can learn a surprising amount about KB simply by playing with a toy, 2D, world. In the first image we set up the world with "molecules" of equal size and mutual attraction, a truly ideal situation. After running the simulation for a while we see the RDFs for g_{11}, g_{22} and g_{12}, all of which are the same. The KBI are also (within the noise) equal. We even find that the chemical potentials are equal. [The chemical potentials are calculated via the magic of Widom insertion. It is not discussed further as it takes us too far from our main goals.]



App 1-3 https://www.stevenabbott.co.uk/practical-solubility/rdf-demo.php

Now re-run it with 2 liking itself (the means are unspecified) 50% more than 1 likes itself and 1 likes 2. Not surprisingly we find a much larger first peak for g_{22} and G_{22} is larger than G_{11} with G_{12} being negative. The chemical potential μ_2 is significantly lower than μ_1 as we should expect.

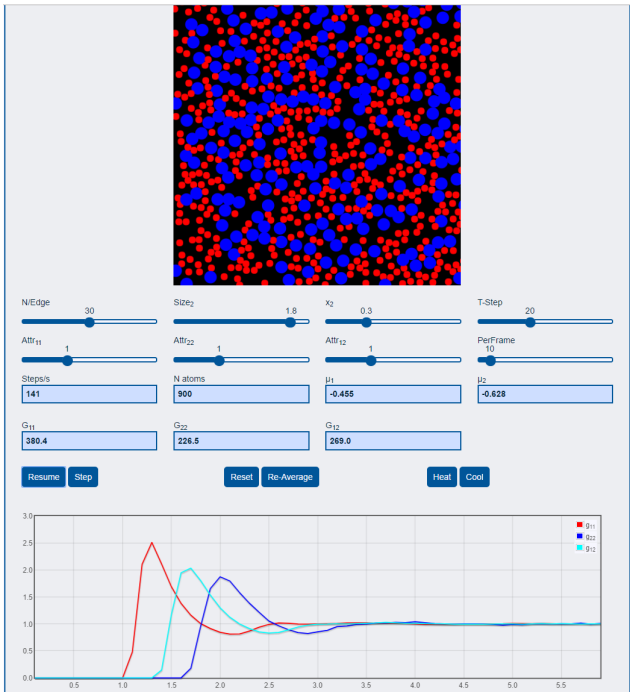
Here we controlled the outcome by specifying that 2 liked itself more. In the real world we would measure the G_{ij} values and be able to deduce that 2 liked itself

6 Unless you are interested in protein stabilisation in deep-water fish. The KB analysis by Shimizu and Smith of the effects of trimethylamine N-oxide requires such a pressure-based analysis.

more than 1. We would then hope to be able to draw some molecular insights from that fact. I have learned so much about KB just playing with such a simple world. I hope you have a reasonably fast laptop so that you can explore many scenarios without getting too bored.

So far, however, the app has not taught us much we didn't already intuit. The next run of the app will, I hope, provide insights that are not so easy to intuit and which are hugely important for all KB, especially work on proteins and other polymers.

1.2.6 Excluded volume



"Excluded volume" is a potent contributor to solubility effects that is at the same time totally simple and (as judged by the literature) massively confusing. It is simple because all it is saying is the obvious - if a molecule is sitting somewhere then you cannot have another molecule in the same place. This trivially obvious fact gives rise to the characteristic oscillations in the radial distribution function, g_{ij} that is integrated to give the KB integrals G_{ij}.

If one molecule is bigger than another then there is more "nothing" before the RDF can start properly. Because the KBI starts integrating right from the start, the more "nothing" the smaller the KBI. We can see this clearly in the same toy KB world as before. Here the relative attractions are all the same. Instead the size of 2 is 1.8, so g_{22} starts much later than g_{11} and g_{12} is in between. The G_{ij} values follow this trend.

This seems trivial, but the outcomes are far from trivial. If you find, one way or another, that a G_{ij} value is lower you can imagine all sorts of reasons for this. But before you invoke, say, water clustering effects, check first whether the explanation is merely that the excluded volume is larger. Decades of speculation have been spent on protein/additive interactions, invoking the magical properties of water when, as it turns out, many of the effects are merely due to excluded volumes. This will be discussed in detail later.

1.2.7 Fluctuation theory

Please feel free to skip this section. I have to add it for completeness and as a challenge to my own ability to understand an important point.

So far I have referred to the theory as Kirkwood-Buff[7], and that is a term that is frequently used. Equally often you hear it described as FTS, Fluctuation Theory of Solutions[8]. Nothing I have written so far has required the idea of "fluctuations", and because I focus so much on KBIs, it makes sense to keep calling the theory KB.

However, there is an equivalent way to describe the calculation of KBIs. Instead of integrating the RDF, it is possible to describe G_{ij} in the following manner, using two equivalent forms. The second form is often seen and is exactly equivalent, but it makes less sense to me in terms of "fluctuations" so I will focus on the first form:

$$G_{ij} = V\frac{\left\langle \partial N_i \partial N_j \right\rangle}{\left\langle N_i \right\rangle \left\langle N_j \right\rangle} - \frac{\delta_{ij}}{c_i} = V\frac{\left\langle N_i N_j \right\rangle - \left\langle N_i \right\rangle \left\langle N_j \right\rangle}{\left\langle N_i \right\rangle \left\langle N_j \right\rangle} - \frac{\delta_{ij}}{c_i}$$

Equ. 1-16

The important bits are neither the V (volume) nor the final term with a Kronecker delta on the top which means it is 0 if i≠j. This is fluctuation theory because of the bits involving N and, especially, the <angle brackets> which mean <averages>. Let us take things one at a time:

$\langle N_i \rangle$ means the average number of particles in the volume being considered. There is nothing difficult about that.

What is δN_i with no angular brackets? It is defined as $N_i - \langle N_i \rangle$ which means in any given part of the solution, the difference between N_i and the average. This is, after all, statistical thermodynamics and we know that around any mean value there is a statistical variation. We also know that some distributions (broad) have plenty of samples with a large variation from the mean and others (sharp) have very small variations from the mean. For KB we cannot use the word "variation" because that has already been used in statistics as the square of the sum of $N_i - \langle N_i \rangle$; it is the familiar term for those who have to calculate, for example, standard deviations in statistics. "Variation", because it is a square is always positive. For KBIs we need the ability to have positive and negative variations from the average. The thermodynamicists have therefore decided to call $N_i - \langle N_i \rangle$ the "fluctuation". If they had asked my opinion as an outsider to their world I

---

7 I know someone who listened to a lecture by Buff. The professor doing the introduction mentioned the "the so-called Kirkwood-Buff theory". Buff corrected him. "No, it is the Kirkwood-Buff theory"

8 An excellent resource for exploring a wide range of KB ideas is the book *Fluctuation Theory of Solutions, Applications in Chemistry, Chemical Engineering, and Biophysics,* edited by PE Smith, E Matteoli, JP O'Connell, CRC Press 2013

would have said "Don't use it; to non-specialists, 'fluctuation' implies a dynamical process. For the specialists it is just a number with no implications of anything actually 'fluctuating'." But it is too late now, we are stuck with the word as applied to KB theory.

These fluctuations, as defined by $N_i - \langle N_i \rangle$ are of no help to us - they don't have any usable meaning on their own. But the numerator of the FTS definition of G_{ij} has angled brackets so we are not using the fluctuations themselves; rather we are using the average of the product of two fluctuations $\delta N_i \delta N_j$. From this definition we can work out two things about G_{ij}:

1. If the distributions of N_i or N_j are very tight, i.e. if there is very little variation from the mean then the numerator is small and G_{ij} is small. The opposite is then true - if there are large variations in the N values (i.e. there are local clusters with very different concentrations) then G_{ij} is large.
2. These fluctuations can be positive or negative, so G_{ij} can be positive or negative.

That is why it is called Fluctuation Theory. If there are big variations, "fluctuations", in concentration throughout the solution, for example by one of the molecules especially liking to be near (or away from) another one, then these give us the large positive or negative G_{ij} values which are signatures of interesting solubility effects. This brings us back to how we approached KB from the start. We look at the RDF around a given molecule and if there is a strong local bias in either direction then we have a large value for G_{ij}. Such a strong local bias means that there is a strong "fluctuation" from the average concentration. So even without those δN_i terms we can say that we understand fluctuation theory, even if we may not find the choice of word especially helpful.

Fluctuations (in the sense of there being regions of different concentration within the solution) are not just inventions of the thermodynamicists. You can see even low levels of fluctuation when you shine an X-ray or neutron beam on the sample and measure the small-angle scattering. Normal dynamic light scattering can see the fluctuations when they are somewhat larger. If they start to approach the limit of fluctuation theory, i.e. just before a phase change, you can even see the fluctuations with your eye, as a "sparkle" in the solution.

That should be the end of this formal digression except for one hint of why fluctuation theory holds out promise beyond the approach I use throughout the rest of this book. There are other fluctuations beyond the concentration-concentration (sometimes called particle-particle) ones that have occupied us. One can also talk about $\langle \delta N_i \delta e_j \rangle$ and $\langle \delta e_i \delta e_j \rangle$, where e stands for energy. As soon as changes in temperature are added to the mix (up to now, everything has been isothermal) we can consider areas of higher- or lower-than-average energy. The concentration-energy and energy-energy fluctuation theories open up some possibilities which, at the time of writing, I am only beginning to grasp.

1.2.8 Scattering

Those with access to small-angle X-ray or neutron scattering can detect "fluctuations" in concentrations directly. One of the many delights of KB is that you can calculate KBI directly from scattering data. This is an important point. There are large amounts of scattering data out there and usually the results are interpreted in whichever language happens to make sense to those doing the study. This means that data from scattering experiments do not tend to build up to a wider understanding. You cannot data-mine for understanding if everyone interprets their scattering according to their own favourite (mis)understanding of solution effects.

Because KB is a universal language with a direct and intuitive interpretation that can be grasped by everyone, it is possible, at least in theory, to reinterpret historical scattering data in KB terms and build up a wide, molecular-based understanding which compares like with like.

Although the equations to convert between KB and scattering are not particularly difficult, it is not something I have ever tried to do, so no app exists to help encourage a mass conversion of scattering data to KB. If anyone with the requisite knowledge of scattering and KB, but without app-creation skills, wants to work with me on such an app, I would be delighted to try.

1.2.9 What have we achieved with all this effort spent on KB?

So far all we have done is to play with some artificial set-ups that allowed us to build our intuitions about what happens at the molecular level when a solute and solvent can interact more or less favourably. Rather than just look at activity coefficients, which are about as deep as we generally go, we have seen that by adding measurements of something as dull as density we can see the extent to which the two molecules form statistical clusters or anti-clusters throughout the full range of mixes. From the KB integrals we can get a feeling, at any concentration, for the excess numbers of molecules (same or different) around each molecule, a fundamental measure from which we can derive many insights into what is going on within the solution.

We have also shown the rather elementary arithmetic going on behind the scenes. The intention was to show how things we know to be important, activity coefficients, are insufficient on their own and that dull things like density and MVol have to be included to allow us to see deeper into what is going on. Future chapters will skip most of the KB details. They are intellectually important but add little to our ability to understand what is going on - provided that the key formulae are made clear (they hold few surprises) and that calculations are done for us via the apps which, of course, they will be.

It can also be admitted that these two-component systems aren't wildly interesting. If you are interested in things like azeotropes then the vapour liquid equilibria and activity coefficients are all you need - the KB integrals do not add massively to the story (though later on I use them to de-mystify some apparently strange solubility phenomena in water). If you are interested in the solubility of a solid in a solvent, the explanations via KB for why something is or is not a good solvent is not in itself of great help. What we need is a method for predicting which solvent is going to be best and it is already clear that KB is great for retrospective understanding but not so good for predicting things.

So the next chapter jumps right in to the complexities of 3-component systems. This is where KB is not just useful, it's transformational. That is why the next chapter is called "Transforming solubility thinking". Before going there, I must issue a disclaimer...

1.3 Other solubility theories are available

In the coming chapters discussions will be restricted to the five principles outlined in the introduction: Ideal solubility (and derivatives), HSP (with related issues such as lattice theory and Flory-Huggins), COSMO-RS, DLVO and KB. The following sections are for those who are interested in knowing about other approaches and my reasons for not discussing them in greater depth.

1.3.1 Abraham parameters

Over many decades, Mike Abraham of UCL has done the hard theoretical and practical work (a rare combination!) to produce a powerful, self-consistent solubility theory based on a LFER, Linear Free Energy Relationship. There are 5 parameters that can be applied to any molecule:

1. E—Excess molar refraction
2. S—Dipolarity/Polarizability
3. A—Hydrogen bonding acidity
4. B—Hydrogen bonding basicity parameters
5. V—McGowan volume

When we come to discuss Hansen Solubility Parameters, we can see that E corresponds roughly to δD, S to δP, A and B are lumped into δH and the essence of V is ever-present within HSP as the MVol. Abraham parameters can also be linked to parameters that are part of the COSMO-RS infrastructure[9] and which find a number of uses beyond COSMO-RS, for example in PSP discussed below.

---

9 Andreas M. Zissimos, Michael H. Abraham,Andreas Klamt, Frank Eckert,and John Wood, *A Comparison between the Two General Sets of Linear Free Energy Descriptors of Abraham and Klamt,* J. Chem. Inf. Comput. Sci. 2002, 42, 1320-1331

With the Abraham approach, if you have a measurement of a property linked to free-energy, such as the solubility of a solute in a solvent, then by a linear free energy relationship you can calculate that solute's solubility in any other solvent. This assumes that somehow you know the parameters of the solute in addition to having access to a list of parameters for standard solvents.

The approach is logical and each time I've seen it applied to a specific problem, it works well. Yet somehow it has not become a general-purpose tool.

1.3.2 MOSCED and PSP

There are two approaches which, like Abraham, use the sound principles of separate parameters for H-bond donors/acceptors: MOSCED (Modified Separation of Cohesive Energy Density) and PSP. MOSCED is briefly discussed in the HSP chapter because at one time, thanks to the inherent superiority of the donor/acceptor approach, it looked an attractive potential alternative to HSP. However it failed to catch on.

PSP (Partial Solubility Parameters), from Costas Panayiotou at U Thessaloniki, is an ingenious development of the solubility parameter approach that also incorporates donor/acceptor ideas and has revised criteria for the equivalents of the Dispersive and Polar components of HSP. By rooting itself in parameters that can be generated directly from COSMO-RS (as mentioned in the context of Abraham parameters) it avoids much of the subjectivity of HSP. It also cleverly distinguishes between three types of solvents:

1. Homosolvated solvents have little or no H-bonding possibilities, e.g. hexane, toluene
2. Heterosolvated solvents do not H-bond internally, but are capable of H-bonding to other solvents, e.g. acetone, acetonitrile.
3. H-bonded solvents that can bond with themselves and heterosolvated solvents, e.g. ethanol, ethylamine.

Unfortunately, at the time of writing PSP have not taken off. My personal experience is that it is rather hard to pin down exactly which of the variants of the PSP should be used within a number of different equations, and the undoubted intellectual advantages of the system have not, in my hands, proven to be substantially and consistently superior to HSP. If they finally catch on it will be a welcome advance in the area of solubility parameters.

1.3.3 UNIFAC

UNIFAC (UNIQUAC Functional-group Activity Coefficients) has a large, dedicated group of users who can rely on it to produce activity coefficients of various mixtures of chemicals at various temperatures. It is, therefore, used extensively in chemical engineering where things like vapour pressures are of

critical importance. It works by breaking each molecule down into its functional groups then adding together all the interactions between those groups in a rather complex manner using a large set of fitting parameters carefully developed over the decades.

I have used UNIFAC on and off over the decades for specific projects, one of which is an app to examine KB parameters. While I appreciate its many strengths, it is certainly not a general-purpose solubility tool of the type that interests me for the purposes of this book.

1.3.4 NRTL-SAC

The non-random two liquid model (NRTL) says explicitly that the behaviour of molecule A around molecule B may well have some specific (non-random) behaviour which will differ from the behaviour of molecule B around molecule A. If one has some coefficients for the NRTL equations then solubility behaviour becomes calculable. For those who are interested in the detailed solubility behaviour of a specific system across a whole range of conditions (e.g. for process engineering) NRTL allows the creation (from a small set of measurements) of thermodynamically plausible fitting parameters that work very well in practice. But this book is more about the prediction and understanding of solubility from molecular structure, and NRTL on its own has little to offer. What it needs is a way to generate fitting parameters from molecular structure.

The SAC (Segment Activity Coefficient) part attempts to do that by assigning three molecular parameters to each molecule: X=Hydrophobicity; Y=Polarity; Z=Hydrophilicity.

The approach certainly has its merits, but the necessary simplifications to create the SAC part, on top of the limitations of NRTL itself mean that the approach isn't wildly successful in the area where it has been targeted: pharma. Because pharma can generally afford the resources, it makes sense to use a more fundamental/powerful approach such as COSMO-RS.

1.3.5 The criteria for a successful solubility theory

The list of theories above is a tiny selection of the more successful ones taken from a vast array of unsuccessful ones, which later I call a babel of languages because they create so much confusion. It is worth taking a moment to consider why a very few succeed and most fail. As a rough sanity check I looked for recent (from 2016) Google Scholar citations of the methods I've described.

The (rounded) results were: COSMO-RS: 1400; HSP:1120; KB:600; Abraham+LFER:200; NRTL-SAC:60; MOSCED:30; PSP:6.

Although NRTL:2200 and UNIFAC:1550 score highly, many of the papers, as mentioned in the context of NRTL, are to do with providing a detailed activity coefficient curve for a specific system. This is of huge importance in chemical engineering but is not the concern of this book.

DLVO (Derjaguin, Landau, Verwey & Overbeek) is in a class of its own with over 3400 hits.

In terms of applicability of the theories, a check of Google Patents for filings since 2014 show HSP: 550; COSMO-RS:200; DLVO:90; NRTL:80; UNIFAC:30 with no mentions of MOSCED or PSP. The HSPiP software is cited 140 times in the same period.

I have created a table summarising my views on the main contenders. This is subjective, but not illogical.

Theory	Assumptions	Usability	Mean field?	Understanding	Prediction
HSP	Many	High	Yes	High	Medium
COSMO-RS	Few	Medium	Yes, but	High	High
DLVO	Many	Low	n/a	Medium	Low
KB	None	High	No	High	Low
Abraham	Many	Low	Yes	Medium	Low
UNIFAC	Many	Medium	Yes	Low	Low
NRTL	Many	Low	~	Low	Low

The "Prediction" column needs some explaining. For a given solute, Abraham, UNIFAC and NRTL can provide excellent predictions, but I have scored them as "low". This is because, in general, you cannot take the excellent solubility predictions from one system and apply them to another. There is little *cumulative* build up of knowledge; each solute starts on its own terms. UNIFAC users will still take exception. However, this book is intended for a broad range of solubility issues and UNIFAC excels in large chemical engineering environments where issues like vapour pressure curves are of far more importance.

The mean field column describes whether the theory assumes the absence of long-range interactions. The "Yes, but" entry in COSMO-RS is there because there are variants such as COSMOmic that can handle non-mean field situations.

We can start our analysis of success and failure with solubility parameters.

The initial version (Hildebrand) used one parameter (dispersion) so this was guaranteed to fail. A version from Blank and Prausnitz used two parameters; this

too was hopeless in the face of reality because alcohols could not be described with the same polar parameter that worked for, say, acetonitrile. Hansen came up with his three parameters, dispersion, polar, H-bonding plus an *extensive* list of values that he and others could use, plus examples from the real world. This created a lot of momentum, and the enthusiastic adoption by industry helped it to become firmly established.

Those who preferred rigour to practicality worked on 4-parameter systems (MOSCED and PSP) and even 5-parameter systems (Karger's chromatography system). The more rigorous systems struggled to create large datasets and link them to the real world, especially a world where solvent blends are the norm. So a system, Hansen, that is intellectually inferior has (so far) won because it is far superior in terms of practical use. My observation is that the predictions of the more complex systems are not so much better than those of HSP to justify their complexity. I assume this is because the errors from the assumptions behind the solubility parameter approach are larger than the refinements from the more complex approaches, so the extra work generates surprisingly little extra benefit. A specific exception where donor/acceptor is obligatory is described in the HSP chapter.

We can imagine that things could have turned out differently. For example, Hoy, working for Union Carbide, created his own version of the three solubility parameter system and Hansen (who knew Hoy) himself acknowledges that it was excellent and gained a lot of traction because of the importance of Union Carbide to the US chemicals industry. And yet this advantage turned out to be a disadvantage. I have been told my someone familiar with the situation that Hoy's approach gave Union Carbide a competitive advantage so he was not encouraged to publicise it or to help it to develop outside. So what may have been a superior system (we shall probably never know) did not, in the end, supplant Hansen's.

Moving on from solubility parameters, those who need greater accuracy for pure systems seemed to have found that the fitted systems such as Abraham, NRTL or even UNIFAC do not give the accuracy that the more powerful COSMO-RS can give without any (user-required) fitting. So the higher cost and steeper learning curve for COSMO-RS are less relevant to the power users who can benefit from the broad range of thermodynamic predictions that it can provide.

So, in practice, we have ended up with two widely-used systems: HSP, a sort of general-purpose off-road vehicle for the rough realities of practical formulations, and COSMO-RS, a sort of S-Class Mercedes ideal for the solubility autobahn.

In addition we need the ideas of ideal solubility to convert crystalline solids into pseudo-liquids to which solubility theories can be applied. We will find that a trivial ideal solubility theory (from Yalkowsky) is often more than good enough

in the face of uncertainties surrounding more profound theories, again showing that "good enough and usable" is better than "superior but unusable".

Then we need DLVO for dispersions. It is a theory universally known to be deeply flawed, yet I have not found any usable alternative. DLVO has more than enough parameters for the formulator to grasp. More complex approaches require even more parameters and most of us have no hope of finding or using them.

And, finally, we need KB because none of the standard solubility theories can cope with the complexities of local ordering/clustering which is so important for issues of solubilization rather than solubility. Of all the theories used in the book, KB has the oddest set of characteristics. First, it is assumption free. Second, KB parameters can usually be determined via very simple experiments, a surprising and transformational aspect of such a profound theory. Third, KB can unambiguously resolve debates that have raged for decades. Fourth, it provides (so far) no predictive power. This fourth aspect is very frustrating, but three out of four isn't too bad.

After that long digression, let us return to the claim that KB can transform solubility thinking.

2 Transforming solubility thinking

The world of solubility science is fragmented into little domains, each of which has its own language and priorities. This fragmentation creates lots of misunderstandings and also wastes a lot of effort in fruitless debates. My reason for writing this book and basing it on KB theory is that there is a glorious unity of solubility science which is most readily analysed and explained using modest variations on the basic KB science we went through in the previous chapter. So KB can transform solubility thinking, solubility language, and our approach to doing measurements in the area of solubility.

This is a grand claim. The first step in justifying it is that KB theory itself is fundamental and assumption-free. It is not some set of clever tricks that happen to do a good job. It is how solubility works.

Against that is the fact that up to recently it has been wildly under-utilised because it is dry thermodynamics full of dull arithmetic and not at all suited for the real world of busy formulators. It was also totally useless until Ben-Naim solved the “inversion” problem that allowed KB integrals to be calculated from experimental data using the methods of the previous chapter. Above all it has been useless because it lacks predictive power as the problem of predicting KB integrals from first principles, e.g. molecular dynamics, has proven intractable.

The reason it is transformational now is that pioneers such as Ben-Naim, Matteoli, Smith, O'Connell, Marcus and Shimizu[10] have shown that KB can resolve with simplicity and clarity many debates that have dogged solubility science for decades. The debates tend to involve water as the solvent, and try to resolve how significant (or not) water structure is to any given solubility issue. Many debates also involve the idea that clustering (in some analogy to surfactant micelles) is helpful for solubilizing many systems. Those readers who have been paying attention may immediately realise why this is mostly wrong!

Rather than discuss abstract reasons why KB is transformational, we will plunge in to a specific area that has caused confusion for many decades: the world of “hydrotropes”.

2.1 Hydrotropes - or are they Solubilizers?

The word “hydrotrope” means many different things to many different people: solubilizer; microemulsion former; solvent; solvosurfactant; nice, friendly, mild surfactant; etc.

---

10 Each of these is an author in *Fluctuation Theory of Solutions, Applications in Chemistry, Chemical Engineering, and Biophysics,* edited by PE Smith, E Matteoli, JP O'Connell, CRC Press 2013. I am grateful for the help from Dr Matteoli in sorting out the KB binary and ternary apps which required techniques far beyond my KB understanding.

In this book I want to use the neutral word “solubilizer” to describe a molecule added in modest quantities that will enable a solvent to dissolve more of a solute. This molecule will generally not be regarded as a solvent in its own right (urea, for example, is a solid), or is added in quantities far less than might be expected if it were acting as a second solvent. This definition leaves many fuzzy edges, but it is far better to use a word “solubilizer” that is relatively obvious and neutral in its meaning than to use a word like “hydrotrope” which has so many meanings and creates more confusion than clarity. Nevertheless, I am forced to use the word hydrotrope because we have terms like "Minimum Hydrotrope Concentration".

We can quickly state the central dogma that has misled hydrotrope researchers for decades, and which KB refutes with simple clarity:

“Many solubilizers are surfactants such as Tweens which form micelles that dissolve the solute. Surfactants have a minimum Critical Micelle Concentration (CMC) below which they have no solubilization effect. Many simple hydrotropes such as urea or nicotinamide have comparable solubilization effects (2-10x increase in solubility) and also show a Minimum Hydrotrope Concentration below which they have no effect. ‘Therefore’ these hydrotropes work via ‘surfactant like’ clusters.”

Although this dogma has been stated in terms of simple hydrotropes like urea , nicotinamide or sodium cumene sulfate, it is regularly used in many other cases with molecules that could be plausibly thought of as surfactant-like and have been sometimes called solvosurfactants.

There has been much earnest work to bolster this logic; all of it fruitless.

In the face of the fact, known to many yet ignored by many others, that urea is totally unable to form clusters in water (its solubility behaviour in water is nearly "ideal"), there is another popular hydrotrope dogma.

“Water is uniquely problematical because it likes to self-cluster. If those clusters can be broken up by a hydrotrope then the water can better dissolve the solute.”

No one doubts that water forms clusters. The “explanations” of how urea or nicotinamide make it de-cluster and increase solubility have always been hard to grasp - and we now know why; because they are wrong. In a later chapter we will mention "chaotropes" and "kosmotropes", terms often used to describe how molecules (urea versus trehalose) or ions (Li^+ versus Cs^+) change the solubility characteristics of, say, proteins. These names imply cause and effect via creating chaos or order in the water structure, yet it turns out that they have contributed little to the real understanding of the effects of the molecules or ions. Let me change that previous sentence to "and have done positive harm to the real understanding of the effects of the molecules or ions". Words like chao/

kosmotrope frame the debate before the debate has even begun. If everyone agrees that you are adding a chaotrope, then the ensuing solubility change must be due to the chaos. These terms subconsciously block off alternative ways to think through the issues. We will find that the effects of chaotropes are highly varied and have *nothing* to do with chaos, but the use of that term can stop us looking for the true explanations. Language matters in solubility.

There is a third popular hypothesis which is backed up by almost no evidence yet has an intuitive appeal.

"Hydrotropes work by forming complexes with the solute".

As we shall see, although this hypothesis is totally wrong, there is a deceptive grain of truth held within it.

The KB work of Shimizu and colleagues (of whom I'm one)[11] shows, with total clarity, how these classic hydrotropes work and the result is a bit of a surprise.

In the basic chapter on KB theory we had molecules 1 and 2. Now we will stick strictly to the convention that 1 is the solvent (in this case, water), 2 is the hydrotrope/solubilizer and u is the solUte. As you can see, it is rather hard to find a numbering system that is clear and helpful as all three start with "sol" - so "u" is used for the solute as it is the first letter that distinguishes it from the others.

We know that the aim is to find the key KB integrals: G_{11}, G_{12}, G_{1u}, G_{22} G_{2u}, G_{uu}, remembering that, say, G_{12} is the same as G_{21}. We also know that we can get all G_{ij} values for 1 and 2 from density and activity (vapour pressure or osmotic pressure) data. Extracting the G_{ui} values requires chemical potential data, which comes from the derivative of the curve of the dependence of solubility on solubilizer, plus an admitted approximation - that at these low concentrations $MVol_u$ is not much different from its nominal value. Finally, we ignore G_{uu} because at the low concentrations of solute it plays no part in the solubilisation equations. As this is KB theory, working all this out is a lot of tedious arithmetic. The paper describes it in detail. In addition, a tutorial review[12] which deliberately only includes one equation, is tied to the apps discussed previously plus the hydrotrope app (below) where we do all the hard work for you. Note that although we have added an approximation in order to calculate G_{ui} values (and omit G_{uu}), this isn't because KB cannot handle the assumption-free case. It's just that to get the exact values we have extra unknowns for which we would need extra experimental data. Our pragmatic judgement is that not only is the extra

---

11 Jonathan J. Booth, Muhiadin Omar, Steven Abbott and Seishi Shimizu, *Hydrotrope accumulation around the drug: the driving force for solubilization and minimum hydrotrope concentration for nicotinamide and urea*, Phys. Chem.Chem.Phys., 2015, 17, 8028

12 Steven Abbott,Jonathan J. Booth and Seishi Shimizu, *Practical molecular thermodynamics for greener solution chemistry*, Green Chem., 2017, 19, 68-75

effort not justified but if it were required, no one would do it anyway as life is far too short and the gain in understanding is going to be trivial.

To anticipate the answer to the core questions about hydrotropes, here is a slogan that works rather well for nearly all solubilizers (including “entrainers” for scCO2): “G_{u2} is good, G_{22} is bad, G_{11} is irrelevant”. The main exception to this rule is the conventional surfactant solubilizer which requires the formation of micelles. How this fits into the broader KB theory, along with solvosurfactants and "pre-ouzo" effects is discussed in the Solubilization chapter. When we come to the aqueous solubility chapter dealing with other effects such as the Hofmeister series, the slogan is even simpler: "If you want to understand u-2 effects, focus on G_{u2}". One reason for writing this book is that until recently the two slogans were largely unknown and had they been known they would not have been believed. Another reason is that the u-2 slogan sounds so obvious that it hardly seems worth stating it. In the world of aqueous solubility this has not been at all obvious, with vast efforts focussed on anything other than u-2.

2.2 G_{u2} is good, G_{22} is bad, G_{11} is irrelevant

With patient measurement of the densities of solutions of the solubilizer, e.g. urea or nicotinamide, in water, along with vapour pressure osmometry data (as "osmolality") measured or found in the literature, G_{11}, G_{12} and G_{22} data could be extracted via the techniques of the previous chapter. The first time we tried this, the density measurements took a few hours because a modern densitometer was available. The osmometry data took considerably longer as the equipment was old fashioned, more or less found in an old cupboard. KB understanding will be seriously hampered until those interested in solubility gain the (modest) funding for good densitometers and osmometers and can integrate them into an HT mindset aided by simple robotics.

The results showed that G_{11} hardly changed with the concentration of urea or nicotinamide. This means that water structure is not changing, so that hypothesis is immediately eliminated. For urea, G_{22} is insignificant (as we said, urea does not self-cluster in water) whereas for nicotinamide it is moderate. Because each can be a good solubilizer, the idea that solubilizers depend on “clustering” via the analogy with surfactants is immediately eliminated.

This leaves us with G_{u2} and G_{u1}. These were derived (again using similar techniques) from an extensive, high quality set of data on solubility versus solubilizer concentration gathered by the team of Prof Gandhi in Chennai. The solutes were relatively simple “drug like” molecules such as methyl benzoate, p-aminobenzoic acid, butyl stearate and ethyl benzene. The Gandhi team investigated several other solubilizers including sodium benzoate and sodium salicylate. With some difficulty we were able to obtain sufficient data to be able to extract G_{ij} data from those solubilizers too. It is striking that the best solubilizer for one solute was not the best for another. So whatever the solubilizer is doing,

there must be specific solubilizer-solute interactions, and the extent of these can be found directly via the KB integrals. The discussion below includes only urea and nicotinamide because they represent rather different types of molecules with very different self-clustering characteristics and very different potentials to interact with other molecules. Readers can go to the app and look at the results for the other solubilizers.

For both urea and nicotinamide with a range of solutes, G_{u1} is insignificant (after all, the solute doesn't like water, which is the problem we are trying to solve with solubilizers) which leaves us with G_{u2} which is significant in all the examples so far studied.

There is one equation that governs the solubilization (and is the only equation in the tutorial review mentioned above):

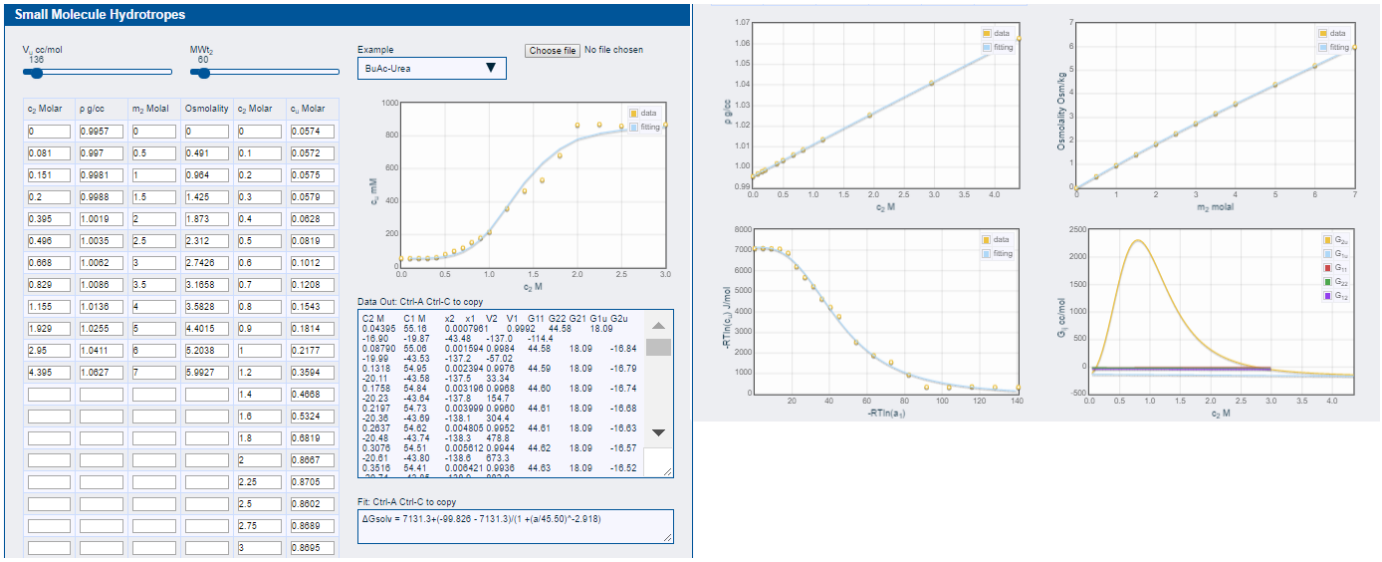
$$\frac{\delta\mu_u}{\delta c_2} = -RT\frac{G_{u2} - G_{u1}}{1 + c_2\left(G_{22} - G_{21}\right)}$$

Equ. 2-1

This is slightly more complex than the equivalent equation for two component systems, but does not introduce any fresh ideas. From our familiarity with the meaning of the KB integrals we can intuit what each of the terms might mean. And the idea of a derivative of the chemical potential holds no special fear - it is just the change of solubility with solubilizer concentration. So we can work out how solubilizers *might* work and from the data can see how they *do* work.

The left hand side is telling us, effectively, how the solubility of the solute u depends on the concentration of the solubilizer 2, with a large negative value being desirable to reduce the chemical potential. With the (justifiable) assumptions that G_{u1} and G_{21} are both unexciting, the right hand gives us the answer: high solubility requires a large G_{u2}, i.e. a strong interaction between the solute and the solubilizer, and a small G_{22}, i.e. a weak self-association of the solubilizer, in direct contradiction to one of the favoured hydrotropy hypotheses.

The app shows everything that is going on. There is the solubility data itself, with the characteristic curve of no effect on solubility till the MHC is reached, then solubility rising to a plateau. There are the raw data on density and osmolality and the fitting functions from which the relevant KBI are extracted. [As discussed earlier, KB calculations depend on *derivatives* of values, not the values themselves. The data are fitted to polynomials so that the derivatives can be calculated at any desired concentration]. Then there are the curves of the KBI, with G_{u2} dominant and G_{22} rather small except (not shown in this image) for nicotinamide which self-associates and, therefore, is less effective than it otherwise would have been.



App 2-1 https://www.stevenabbott.co.uk/practical-solubility/kb-hydrotropes.php

I remember clearly my excitement when I first understood the equation explaining hydrotropy because it was followed almost immediately by confusion. The G_{u2} curve shows a large maximum, which is associated with the increase in solubility but then it reverts to 0 at high solubilizer concentrations. Doesn't a G_{u2} of 0 imply low solubility? Yet although the solubility has plateaued, it hasn't become lower. This is where it is important to remember that G_{ij} values are integrals with respect to the average bulk solution. Suppose the solute needs, on average, 1 solubilizer next to it in order to feel happy in the water. When the levels of solubilizer are low there is no chance of attaining that average unless there is a high G_{u2}. But at high concentrations of solubilizer just the average number of solubilizer molecules in the neighbourhood is enough to keep the solute happy, so there is no need for a high G_{u2} value to keep the solute in solution.

The loose way of describing the situation is a necessary way for building an intuition about another explanation sometimes offered for solubilizer effects. I chose 1 solubilizer deliberately in the previous paragraph because the false idea that “1+1= a complex” is seized upon by those who favour the “solubilizers as complexing agents” explanation - despite the fact that unambiguous evidence for complexes had been hard to find. Thermodynamically there is a big difference between the KB behaviour of a complex and the general shape of the KB curves found for all the solubilizer cases studied so far.

It is important to remember that KB curves are statistical constructs. With plausible assumptions we can calculate excess numbers, N_{u2} of solubilizer around the solute. These numbers are non-integers and vary smoothly from 0 to a maximum and back to 0. The maximum may be 0.789 or 1.234 or 2.345. This is nothing to do with “complexes”. And remember the warning in the discussion of excess numbers; it is far from clear what the “ideal” case should be that defines the excess. The general shapes and magnitudes of excess numbers

are highly insightful but must be never taken too seriously, as the same KB data in different hands might yield equally plausible, but different, excess numbers. Whatever the exact numbers and shapes, it is clear that even if there is a headline number of 1 excess molecule this is *very* different from a 1:1 complex, and 2 excess molecules are definitely not a 1:2 complex. It is hard to imagine how such a complex could be produced via the generally dull solubilizers and uninteresting solutes used in the experiments.

So with the solubility data and some density and osmometry data a debate that has raged for decades has been definitively settled in favour of an explanation that was hardly ever raised. As we shall see, the same approach immediately settles a debate in a very different area.

But there is one thing missing from the explanation. Why is there a Minimum Hydrotrope Concentration? The failed explanation (hydrotrope clustering) that created an analogy to CMC was attractive because it offered an apparent explanation for the MHC, even though there was never any good evidence for why the MHC was at concentrations of ~1M rather than the μM or mM concentrations of surfactants.

The real explanation required some virtuoso KB theory from Shimizu and Matubayasi[13] that is too complex to describe here. Although the theory is complex, the explanation is via another KB integral that poses no extra difficulty, though admittedly it takes some getting used to.

The MHC kicks in when $G_{u,22}$ becomes large. What does this mean? We know that G_{22} is the self-association of the solubilizer and we also know that this is a bad thing in terms of solubilization. We know that G_{u2} is association of the solubilizer with the solute which is the driving force of the solubilization. But G_{u2} cannot give a MHC. $G_{u,22}$ is the self-association of the solubilizer *induced by the solute*. This type of self-association has, to the best of my knowledge, never been suspected before, which is why MHC had always been a mystery. It cannot kick in until there is enough solubilizer available to self-associate, but once there is enough, the self-association around the solute creates a significant driving force for further solubilization. It is fascinating that self-association is bad (because it effectively removes solubilizer from the solution) yet when the solute itself induces self-association of the solubilizer, good things happen. Remember, too, that this effect works for urea, so there is solute-induced urea-urea clustering even though urea does not self-cluster.

---

13 S. Shimizu and N. Matubayasi, *Hydrotropy: Monomer−Micelle Equilibrium and Minimum Hydrotrope Concentration*, J. Phys. Chem. B, 2014, 118, 10515–10524

For those interested in exploring the MHC, a complementary theory which is less exact but is much easier to grasp has been created[14] by the same authors as the full MHC theory and the calculations (3 parameters to generate the fitting curve to the solubility data) are included in the app.

With a complete explanation of classic small-molecule hydrotropy is that the end of the matter? No! Like everything else involving KB the explanation is purely retrospective - we can analyse why, say, urea is better than, say, nicotinamide for solubilizing methyl benzoate because although nicotinamide has a higher G_{u2} than urea it also has a bigger self-association, G_{22} which, on balance, reduces the solubilization. What we *cannot* do is predict in advance which would be the better solubilizer. To do this we would need to be able to calculate G_{u2} and G_{22} from first principles using, say, molecular dynamics. But the subtle balance of many effects within such calculations do not currently make them a reliable way to form predictions.

Part of the problem is that with no consistent science for investigating the various solubilizer effects, there has not been, until recently, a set of test cases or datasets for those armed with sophisticated predictive tools to validate or optimise their systems.

A good example is that there has been a lot of SAXS/SANS work on systems that may or may not be good solubilizers. The data can be, and has been, used to calculate KB integrals. But those, for example, who assumed that self-association was a necessary attribute for hydrotropy would have been focussing on the wrong measurements. Strangely, getting funding for such sophisticated experiments is probably easier than getting funding for very dull density or osmometry measurements, though these latter techniques provide a lot of insights for very little work if the equipment is modern.

One reason for writing this book is to encourage a change in the attitudes to research on solubilizers. With some good sets of high-quality data and the clarity of the challenge of being able to predict the balance of KB integrals, theoreticians will be able to rise to the predictive challenge. Happily, a 2017 paper[15] has shown that this sort of approach is possible. For a single solute, indomethacin, a large number of hydrotropes were used. The data were fed into an artificial neural network and a good predictability with respect to the training set was obtained. By combining the artificial neural network with some human intelligence, it was possible to identify a new (i.e. not in the training set) potential hydrotrope with predicted high efficacy. When tested, the new molecule proved to be highly successful.

---

14 Seishi Shimizu and Nobuyuki Matubayasi, *The origin of cooperative solubilization by hydrotropes*, Phys. Chem.Chem.Phys., 2016, 18, 25621

15 Safa A. Damiati, Luigi G. Martini, Norman W. Smith, Jayne M. Lawrence, David J. Barlow, *Application of machine learning in prediction of hydrotrope-enhanced solubilisation of indomethacin*, International Journal of Pharmaceutics 530 (2017) 99–106

Hopes for a widespread adoption of this KB-driven approach would be worthless if it helped only with small-molecule solubilizers and if the goal was simply an academic understanding of a fascinating problem. In the Solubilizers chapter the wider need for more KB understanding within industry is explored. At this stage of the book it is important to justify the "transforming solubility" title by looking at a seemingly different solubility challenge that has defied explanation before KB was used. The example is very much real-world, with major economic impacts in terms of green (or not) extraction of natural products.

2.3 scCO2 entrainment

There is an obvious attraction for using CO2 as a solvent - at the end of the process the CO2 can simply evaporate away, leaving the solute behind. The classic use for this is the decaffeination of coffee, with CO2 replacing solvents such as dichloromethane.

In two ways, however, CO2 is a poor solvent. First, the molecule shows few features that make it want to interact with most solutes of interest - certainly not caffeine. Second, ignoring frigid liquid CO2 it only becomes dense enough to show significant solubility properties when it is brought to something like 30°C and 120bar pressure when it becomes a supercritical fluid, scCO2. Although much is made of the supercritical state, in solubility terms it really isn't very interesting. Yes, viscosities are low so the kinetics of dissolution can be faster than with conventional solvents. But apart from that, to the solute it is a relatively low density fluid, which means that the in addition to CO2 being chemically uninteresting, the density of molecular interactions is low, making it an even worse solvent. It is commonly said to act like iso-pentane, not a solvent that most of us would choose to use for most solutes.

So the number of systems for which scCO2 is a good, practical solvent is rather small. The inconvenience of using a high-pressure system then makes it even less attractive to use scCO2. And despite the claims of it being a "green" solvent, the energy cost of creating the supercritical fluid and the engineering cost of a high-pressure system greatly dilute its claims to help save the planet.

That would be the end of the story if it were not for "entrainers". It turns out that if you add a few % ethanol, acetone, ethyl acetate etc. to an scCO2 system you can get some reasonable solubilities. So caffeine is barely soluble in scCO2 but adequately soluble in the presence of a few % ethanol (and also in the presence of water, but water is very insoluble in scCO2 so its use as an entrainer is limited). The literature is full of examples of the use of small % of ethanol or acetone to get most of the advantages of scCO2 without the drawback of minimal solubility. The question then arises as to how the few % of entrainers can yield a significant increase in solubility, and how to find the optimum entrainer.

The literature is very confused about this, partly because the scCO2 system itself is complex with its strong dependence on temperature and pressure. To fully understand what is going on you don't just have to measure solubilities but also the density of the system. For conventional solvents, there is no tradition of worrying about densities as they don't change all that much. As mentioned in the previous section, this is unfortunate because measuring densities provides lots of KB insights. For scCO2 it has always been clear that the solute and the entrainers on their own have a large effect on the density of the system (after all you are adding a solid and a liquid to a weird supercritical state), so it has been seen as necessary to measure the densities of the scCO2 system. And to gain understanding, it has been traditional to measure solubilities at a range of temperatures and pressures.

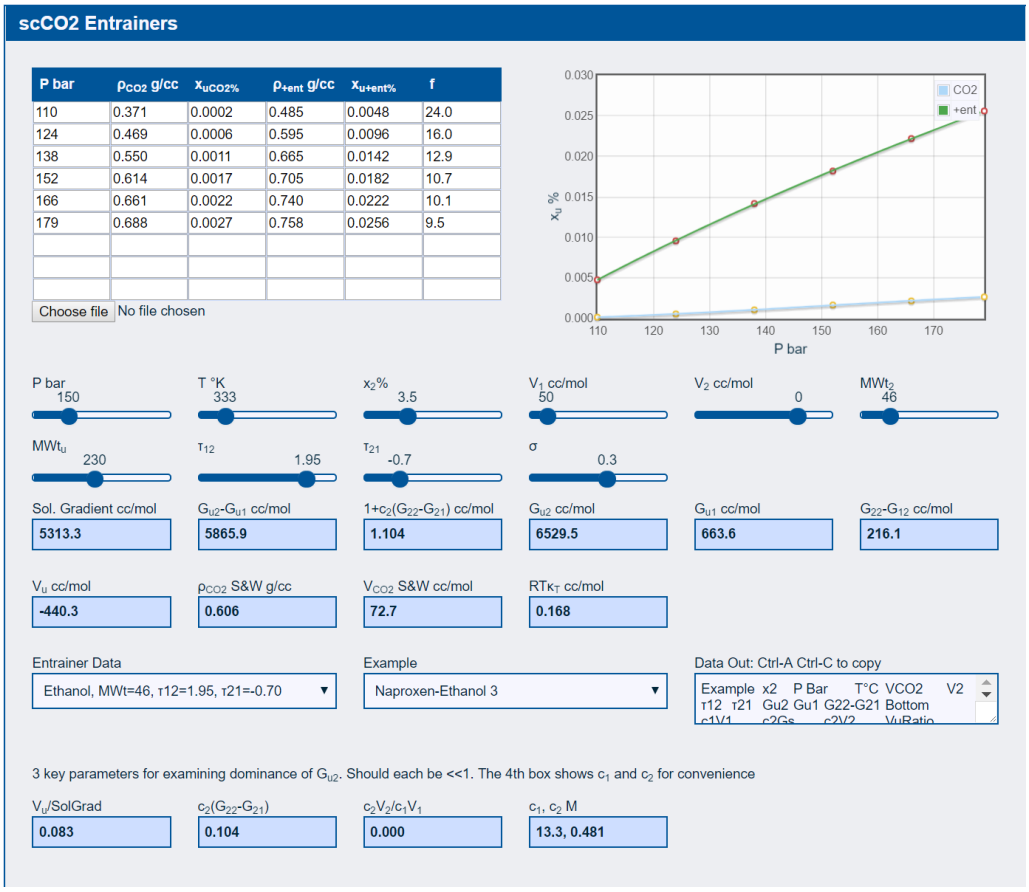
The problem is that these datasets have then been analysed via whatever ad hoc approach appealed to the researchers. This would usually involve some EoS, Equation of State, it might involve some NRTL (Non-Random, Two Liquid) activity coefficient estimates, and then some fitting to some plausible equation. The net result has been a large number of papers that have failed to build a broad understanding of what is going on. How can you compare some fitting of one EoS or some ad hoc formula to a different EoS or to another ad hoc formula? You cannot.

Once more, KB is able to unify a complex problem with elegant simplicity. A paper by Shimizu and Abbott[16] analysed the entrainer effect on 16 systems. The outcome is the same as the hydrotrope story. With 1 being the solvent, scCO2, 2 being the entrainer and u being the solUte it was possible to look at how entrainer and/or solute affected the CO2 self-association (G_{11}), which is hardly at all. It was also possible to measure the amount of entrainer self-association (G_{22}) and see its effect on solubility of the solvent, which was small and negative. Not surprisingly, the large effect, in all 16 systems, arose from G_{u2}, i.e. from entrainer solute interactions. It is equally no surprise that the entrainer effects were stronger at lower pressures when the CO2 was least dense and therefore a poorer solvent.

These conclusions were robust even when it was not possible (through lack of experimental data in the original literature) to calculate all the required parameters. For example, $MVol_u$ and $MVol_2$ are *not* equal to their "normal" values and depend strongly on pressure and temperature. In some cases their values are *negative*, meaning that they take up less space than if they weren't there, i.e. they help to compact the overall system.

---

16 Seishi Shimizu and Steven Abbott, *How Entrainers Enhance Solubility in Supercritical Carbon Dioxide*, J. Phys. Chem. B 2016, 120, 3713−3723



App 2-2 https://www.stevenabbott.co.uk/practical-solubility/scco2.php

To make it possible to write a convincing academic paper even though there were so many uncertainties due to lack of data, the paper was linked to the scCO2 app, along with an invitation to inspect, criticise and improve the open source code or the theory. For those who simply want to check the robustness of the conclusions, values for uncertain parameters such as the MVols can be changed with a slider and the impact of the changes on the overall conclusion can be inspected. Happily, no reasonable change to any of these values makes a significant change to the overall conclusion that the solubility of the solute is enhanced via specific entrainer-solute interactions (G_{u2}) and reduced by any tendency of the entrainer to cluster (G_{22}).

Just as with the hydrotrope/solubilizer story, the definitive conclusions from the KB analysis are just a beginning. By having a universal, assumption-free method for looking at the complex interactions within the scCO2 system it is possible to build a clear picture of what is happening with solutes, entrainers, temperature and pressure. We can compare datasets objectively, unlike all those papers that perform analyses via ad hoc EoS approaches. But KB does not allow us to *predict* what will be the effect of any specific entrainer on any specific solute. For this we need other solubility tools, coupled to a large database of entrainer data so we can compare predictions to reality.

As a contribution to this process, Abbott and Shimizu[17] extracted a large number of so-called k_{av} parameters from the confusion of measures of the effectiveness of entrainers within different papers, many of which do not supply enough information to perform a proper KB analysis. The simple formula used says that the solubility S in the presence of entrainer is given by $S=S_0(1+k_{av}y_2)$ where S_0 is the entrainer-less solubility and y_2 is the mole fraction (typically 1-3%) of the entrainer. These k_{av} values are comparable unlike the frequently reported "enhancement factor", f, which is the ratio of solubility with and without entrainer. Values of f cannot be meaningfully compared directly because a larger f might come from a larger % entrainer rather than a stronger intrinsic entrainment capability. A large k_{av} means that the effect of the entrainer is large. For those who are familiar with the topic, the equation is a linearised Sovová formula where the linearisation loses some accuracy but gains comparability. Although these k_{av} values are approximate, they are comparable across solutes and entrainers and allow the big picture to be seen. They are also directly linkable to G_{u2} values, giving them some fundamental value. Sadly, the message is that most entrainers for most solutes are not especially useful and only solute-entrainer pairs with strong H-bonding potential give significant solubilization effects. As you can explore for yourself in the app (by looking at the enhancement factor, "f", column) naproxen, for example, gains a ~10x increase in solubility with ethanol (which can be a donor/acceptor with the naproxen acid and ether groups) and ~5x increase with acetone which can only be an acceptor to the acid group. With benzoic acid, the enhancement factors for ethanol, acetone and hexane are 4, 2 and1.5. For solutes such as 1,8-cineole with just one ether group, the enhancement factors (not included in the app) for ethanol, acetone and hexane are each ~1, i.e. they have no effect.

Note that almost no one[18] in the scCO2 literature had ever asked what the relative 1-1, 2-2, u-2 (etc) attractions were, so all sorts of well-meaning but data-less explanations for, and equations for, entrainers have appeared in the literature. As soon as it is clear that u-2 is the only thing that really matters, the question of what would promote large u-2 values can be posed and the rather obvious answer "H-bond interactions" can be explored. To do the exploration, the 100+ datasets on scCO2 entrainers had to be re-analysed in terms of a common, simple formula (the k_{av} formula) that lost some precision and gained a lot of clarity. Just a glance at the data shows that the H-bond hypothesis has a lot going for it. One can then attempt to falsify the hypothesis by looking for data on other potential H-bonding entrainers such as acetic acid and ethyl lactate. They did not refute the hypothesis, and the strong effect in a single example of ethyl lactate as an entrainer raises the question of why this green(ish) entrainer is not used more often. Water is a problem because it can only be added at

---

17 Steven Abbott and Seishi Shimizu, *Understanding entrainer effects in scCO2*, in Thomas M Attard and Andrew J Hunt (Editors), *Supercritical and Other High-Pressure Solvent Systems*; Green Chemistry Series, Royal Society of Chemistry, 2018.

18 The use of FT-IR to identify solute-entrainer interactions is discussed in the book chapter

relatively low concentrations. Even so, the evidence is borderline (i.e. water is a poorer entrainer than expected), certainly not strongly supportive. But we know that G_{22} for water is likely to be large (water is, after all, not very soluble in scCO2) so some careful experiments to measure G_{22} and see if it is high enough to explain the relatively low efficacy of water would be a good test of the hypothesis.

I hope that the preceding paragraph sounds totally obvious. Yet decades of scCO2 experimentation had failed to come up with any falsifiable hypotheses (though the H-bonding idea was frequently mentioned). The scCO2 world were persistently asking the wrong questions and were persistently satisfied by ad hoc calculations against whatever model most appealed to the specific researchers. That is not good science and is especially bad in an area which is supposed to be green. Research resources are precious (and polluting) so they should be used with care, which means using the right language, asking the right questions and finding falsifiable hypotheses. In the book chapter we claim that a few months of focussed work on a small number of carefully-selected solutes with a small number of carefully-selected entrainers, using smart, high throughput techniques would basically resolve the scCO2 situation once and for all. That is a few months of work compared to decades that have achieved remarkably little. That is what I mean about transforming solubility thinking.

Before discussing HSP, COSMO-RS and DLVO, let us explore some other ways that KB is revolutionising solubility thinking.

2.4 KB and proteins

It has long been known that the addition of small molecules such as urea or guanine, or larger molecules such as sugars or polyethylene oxide (PEO), to a protein solution can change the protein's conformation, stability, self-association, gelation and so forth. Because this is taking place in water, the old favourite, water structure, has been invoked many times and provided no lasting understanding of what this actually means. An alternative explanation has invoked binding of the salts or sugars to the protein, again with little insight into what is actually happening.

The KB approach[19] once again introduces clarity to a confused situation. The specific details of the paper are discussed in the Hofmeister section of the Aqueous Solubility chapter. Briefly, using 1 for water, 2 for the modifying molecule and u for protein we have the usual choice of effects based on G_{11}, G_{22}, G_{u1}, and G_{u2}. As in all other cases, the water structure G_{11} is irrelevant. Usually G_{u1} is also small or irrelevant. In most cases, it is G_{u2} that controls the effect - in other words the effect of the additive is due (for better or worse) to the

19 Seishi Shimizua, William M. McLaren and Nobuyuki Matubayasi, *The Hofmeister series and protein-salt interactions*, J. Chem. Physics, 124, 234905 (1-4) 2006

interactions between the additive and the protein, rather than via anything to do with water.

The temptation, then, is to look for interesting reasons for these interactions, and in the case of the Hofmeister ions this is entirely justified. But often it is a supremely uninteresting interaction that causes the effects, the excluded volume. We have already met the trivial concept behind it - where one molecule is, no other can be in the same place. This means that the RDF gets off to a bad start, with a zero value over a large radius r and that in turn means that the integral that creates G_{ij} is off to a negative start because the integral is RDF-1, where 1 is the long-term average. Two large molecules will have a large excluded volume so even if their general interactions are neutral, G_{ij} might be negative. And that is what happens when a molecule of, say, sucrose is added to a protein solution. It might be entirely neutral in its interactions, but the excluded volume effect means that G_{u2} ends up being negative, resulting in a stabilisation of the protein conformation (pushing it towards a folded state) or the creation of gelling (pushing too hard). When it comes to proteins it is always tempting to come up with exciting explanations for various effects, but unless the protein and an additive like a sugar have some specific reasons for interacting strongly, the sugar affects the system purely through excluded volume.

Against this background, any sugar which happens to solubilize a protein solution (tendency towards unfolding or reducing gelation) and which happens to be the same general shape/size as a sugar which encourages folding *must* have some specific positive interaction. For those who are interested both in the dull general effect and the occasional specific effect (sugars with equatorial -OH groups), there is an excellent KB analysis of the effect of sugars on the food thickening agent carrageenan by Shimizu and co-workers[20].

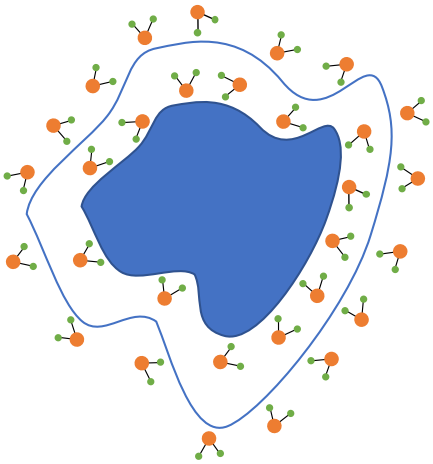
2.5 The problem with the RDF

There is a problem with the RDF which is especially acute with irregularly-shaped proteins: what do we mean by distance between two molecules? It is especially hard to derive a meaningful interpretation if the distance is defined as being between the centre of the (large, complex) solute (such as a protein) and a medium-sized molecule (such as a sugar). I must stress that this is *not* a problem for KB itself, which is assumption-free. It *is* a problem for how we interpret KBI and for how we derive meaning from molecular dynamics simulations.

---

20 Richard Stenner, Nobuyuki Matubayasi, Seishi Shimizu, *Gelation of carrageenan: Effects of sugars and polyols*, Food Hydrocolloids 54 (2016) 284-292

A paper by Martínez and Shimizu[21] uses the examples of urea and TMAO (trimethylamine N-oxide, discussed in detail later) interacting with Ribonuclease T1 as computed via molecular dynamics. From the simulation the RDF can be calculated as normal, and the results provide no obvious insights into why urea prefers to be with the protein and the TMAO is excluded, facts that are obtained from the experimental data by the normal means. Using a new "minimum distance" RDF, not only does the integration for the KBI become much easier (it converges quickly) but the RDF makes intuitive sense.



The trick is to define the RDF on the basis of the minimum distance between an atom on the protein and an atom on the relevant solvent (in this image water) or added molecule (urea or TMNO). The irregular blue shape is the protein and a selection of water molecules close to it each have an obvious closest distance which is included in the RDF count. The fact that the shape is irregular means replacing the πr^2 term in the integration for the KBI with a specific shape term (the surface defined by the minimum-distance r to any solute atom). That shape term is the original shape for the inmost set of molecules, then something like the outline shape shown further out, shown with a new set of molecules with their individual shortest distances. Convergence with radius is rapid compared to the generic KBI - in this example a distance of 8 Å does a good job so the calculations become more tractable. The "nearest-atom" definition ends up giving atomic, rather than molecular, densities. This also turns out to be convenient because atom-specific RDF are deeply meaningful in the new procedure. Conversion to real-world densities (and KBIs) requires a normalization via the concentration of minimum distances in the absence of solute-solvent interactions, which is used instead of the bulk concentration that appears in standard RDFs.

The reason that the atom-based approach is so attractive is that the new RDFs peak at meaningful distances, especially H-bond distances. The water-protein RDF shows a strong peak at the expected H-bond distance ~1.8 Å, as we would expect, plus a "second shell" peak at ~2.7 Å. There is a similar peak for urea and the protein, which is consistent with urea having favourable overall interactions with the protein. You can also find TMAO peaks near the protein, which does not fit with the fact that the TMAO interaction is overall unfavourable. This is, again, why KB analysis is so powerful. It is not at all surprising that TMAO has at least *some* H-bond interactions with the protein, and that is what the RDF shows. The important point is that these interactions are relatively

---

21 Leandro Martínez and Seishi Shimizu, *Molecular Interpretation of Preferential Interactions in Protein Solvation: A Solvent-Shell Perspective by Means of Minimum-Distance Distribution Functions*, J. Chem. Theory Comput. 2017, 13, 6358−6372

smaller than the urea ones and the overall effect is that the large excluded volume term from the protein is not overcome by those interactions.

It seems to me that this more subtle way to analyse MD calculations, especially for the complex interactions with large molecules such as proteins will prove to be a powerful, general-purpose tool for understanding what is going on.

2.6 KB and Ionic Liquids

In the Green Solubility chapter I discuss the rise and fall of the myth that ionic liquids (ILs) are green and wonderful. It is now clear that they will almost never find mass use as green solvents. Instead they will be used on solubility problems for which other classes of solvent are intrinsically useless. The importance of ILs will not be as replacements for conventional solvents but as solvents that can do what cannot be done by conventional solvents.

Because their solubility properties are unique it is important to identify what aspects of solubility make them unique. The first of these is the strong dependence of their properties (e.g. solubility and viscosity) on water content. It is easy to speculate about what happens between the water and the ions. With sophisticated equipment it is possible to look at various ion-ion and ion-water interactions. The beauty of KB is that the simplest basic experiments (densities, water vapour pressure ...) provide the assumption-free answers to key questions, via the G_{ij} values where 1=water and 2=IL. The answers to 3 questions for 3 aprotic and 3 protic ILs are provided in a paper[22] by Shimizu and colleagues. First, the values of G_{22} do not change much with water content, so the water has little effect on ion-ion interactions, whether the ILs are protic or aprotic. Second, from G_{11} values, there are very large differences between ILs on the degree of water self-association. At low water concentrations, the self-association varies from very low (i.e. water-IL interactions are strong) to very high. And at higher water concentrations (above 0.1 mole fraction) many ILs show no interesting effects but a very hydrophobic IL shows a huge increase of G_{11} indicating movement towards phase separation. Thirdly, the ion-water interactions via G_{21} show big differences in dependence on the mole fraction of water. The data go up to 0.5 mole fraction which sounds a lot, but with a typical 10x difference in MWt, that is only ~10 wt% water. The most dramatic changes in G_{ij} values happen between 0 and 0.1 mole fraction, i.e. ~1 wt% water which (sad experience has shown) can easily contaminate a nominally "dry" IL.

The precise details are not important as the six molecules in the study are an insignificant fraction of ILs. The important point is that just these six are sufficient to show that for any IL question of interest (e.g. protic versus aprotic) there are no simple, general rules of how water affects the IL so there can be

---

22 Joshua E. S. J. Reid, Richard J. Gammons,John M. Slattery,Adam J. Walker, and Seishi Shimizu, *Interactions in Water−Ionic Liquid Mixtures: Comparing Protic and Aprotic Systems*, J. Phys. Chem. B, 2017, 121 (3), 599–609

no simple general rules of how water in ILs will affect their behaviour. For any given IL system, the nature of the water-IL interactions could be worked out rather indirectly from extensive experiments with, say, NMR and FT-IR. The message here is the same as in other cases: simple experiments directly reveal fundamental information. On the basis of these fundamental insights it can be highly fruitful to use specific techniques such as FT-IR to gain more details. The later section on Hofmeister ions shows that such a parallel approach works very well.

Those who wish to go beyond the limitations of KBI to look for predictive capabilities could readily get a robot to do the large number of simple high throughput experiments required to get enough data to be able to mine it for scientific insights. This brings us to an interesting point.

2.7 KB and the end of stamp collecting

The literature in the areas discussed in this chapter seems to me to be filled with "stamp collecting" papers. Some random data on some random system is analysed in an ad hoc manner and the paper adds no significant advance in understanding or predictability to the relevant community. Although one might argue that science ultimately can advance via such an approach, it can equally be argued that this is a deeply inefficient way.

The better way is to have a clear hypothesis which, if true would underpin future work or, if false, would allow other hypotheses to have a chance. The trouble with many of the topics in this chapter is that hypotheses, if they existed at all, were generally too vague ("x is caused by changes in water structure") to be refutable.

A defence of stamp collecting papers is that when there are enough of them, they can become "big data" and be "mined" for information which might, in turn, generate clear hypotheses to be tested. As I have spent many deeply frustrating days trying to mine data from many such papers I can assure readers that it takes only one nugget of information to be missing from a paper to make it impossible to mine anything useful. This is in addition to the frustration of having to read data off graphs (an on-line tool for doing this is on the Bookmark bar in my browser as I have to use it so often) and to cope with strange and inconsistent units.

We cannot do anything about the past. But we can do something about the future. Imagine what would happen if all papers about these complex solubility issues routinely provided relevant G_{ij} values. We would have the chance to see a bigger picture, to support/refute hypotheses and to start generating prediction rather than retrospection. If obtaining these data required exotic experimental techniques then such a dream would be fatuous. But because it requires low-

tech measurements on solutions that have to be created for the other aspects of the research, we gain a lot of information for very little extra work.

My personal commitment is to find ways to crank out large volumes of KBIs in rather boring systems so that people smarter than myself can data-mine them for deep, predictive insights. This is a task that would be career suicide for an academic. And a decade ago it would have been a task that could only have been done by a well-funded academic or industrial organisation.

Fortunately, this is the 21st century and it is now cheap and easy to create sophisticated chembots and mixbots with versatile browser-based interfaces for data gathering and analysis. The Makers Movement and Open Source have made these things so simple that even I can do it.

My first chembot took me a weekend to construct, and I'm a complete klutz. When I took it to a university lab to get it up and running, it was all done in a few hours. The fatal flaw in the plan was that although the chembot could easily make 100 ternary solvent blends (yes, very boring), getting the density and vapour phase data from those 100 tubes was not trivial, so the approach was never going to scale. For example, caps had to be screwed on to the 100 tubes by hand then unscrewed ready for measurements in a densitometer which in turn required taking out samples with a syringe and injecting them into the densitometer. The tasks are trivial, but they take up a lot of time and are very tedious. Automating them is not something that I can engineer in a weekend.

It was my fault for not thinking things through. At the time of writing, a second-generation mixbot which allows density and vapour phase data to be gathered immediately after mixing is being tested. When this works, a general-purpose KBI machine will be able to crank out thousands of values on relatively simple systems (hydrotropes, solubilizers, ternary solvent blends), all at trivial cost if you happen, as I do, to have some brilliant colleagues who are smart enough to invent the key element of the mixbot and Open Source it to the wider community, and who can work out how to re-purpose an old mass-spec system lying around doing very little.

3 Hansen Solubility Parameters

Solubility is so obviously complex that it seems unlikely that any simple theory can be of much use. The assumption-free KB theory isn't especially hard, but it is still rather abstract and it does not (yet) provide us with any way to go from molecular knowledge to solubility knowledge. In principle, as discussed in the KB chapter, we can get a lot of information from molecular dynamics, but in practice MD has not proved to have the right balance of speed and simplicity.

So decades ago when Joel Hildebrand proposed that a lot of solubility issues could be captured in a single number, the Solubility Parameter, SP, there was both scepticism and relief. Scepticism from experts that such a simplistic approach could be of any value and relief from formulators who could see the possibility of a simple and respectable route to solubility predictions.

Both attitudes were somewhat justified. For the relatively simple systems that Hildebrand was analysing, the theory proved to be not at all bad, though its limitations were clear to everyone with the knowledge to comment on it, justifying the sceptics. The problem for the enthusiasts was that Hildebrand had proposed his theory only for molecules that had no significant polar or H-bonding possibilities, and that the enthusiasts ignored this severe restriction. So it was easy to show theoretically that Hildebrand SP were utterly inadequate for the real world, and it was equally easy to be a disappointed formulator when SP predictions turned out to be misleading.

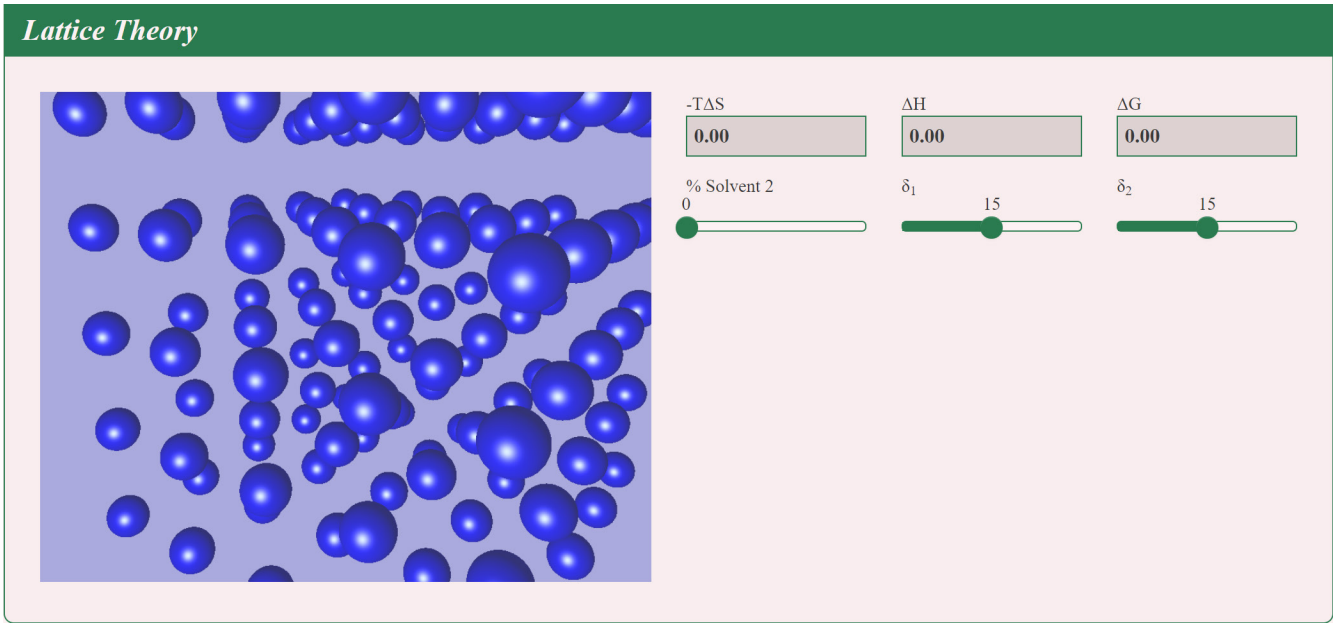
It is a matter of great wonder to me that in 2017 people are *still* using Hildebrand SP in situations where they are refuted by the theory's own assumptions. So here is the first take-home message on SP: Never, ever, use Hildebrand SP; they were, are and always will be utterly useless for most real-world formulation issues. This denunciation of these SP is in no disrespect to Hildebrand. His basic insights into SP revolutionised practical solubility thinking and when the hard work was done (by Hansen) to overcome their intrinsic limitations so that they worked for polar and H-bonding systems, Hildebrand's vision became a practical reality. So this chapter is about the hugely (unreasonably) successful Hansen Solubility Parameters (HSP - note they can be singular or plural) approach.

But before getting to HSP, let us first explore the toy world that Hildebrand created in order to grasp what SP are about.

3.1 Regular Solution Theory and Lattice Theory

Imagine a world where all solvent molecules are spheres that, conveniently for us, sit on a lattice structure. Or, instead, don't just imagine it, have a look at it in 3D:

Here we see a set of blue molecules (call them 1) neatly arranged on a cubic grid. The fact that real molecules are not spherical, nor do they sit on a regular grid does not bother us. Our concern here is that the molecules have an attraction for each other which gives them an enthalpy and they have an ordering which gives them some entropy.



App 3-1 https://www.stevenabbott.co.uk/practical-solubility/lattice.php

We have no interest in the absolute enthalpy or entropy. What concerns us is what happens when we replace some blue molecules with red ones (of type 2).

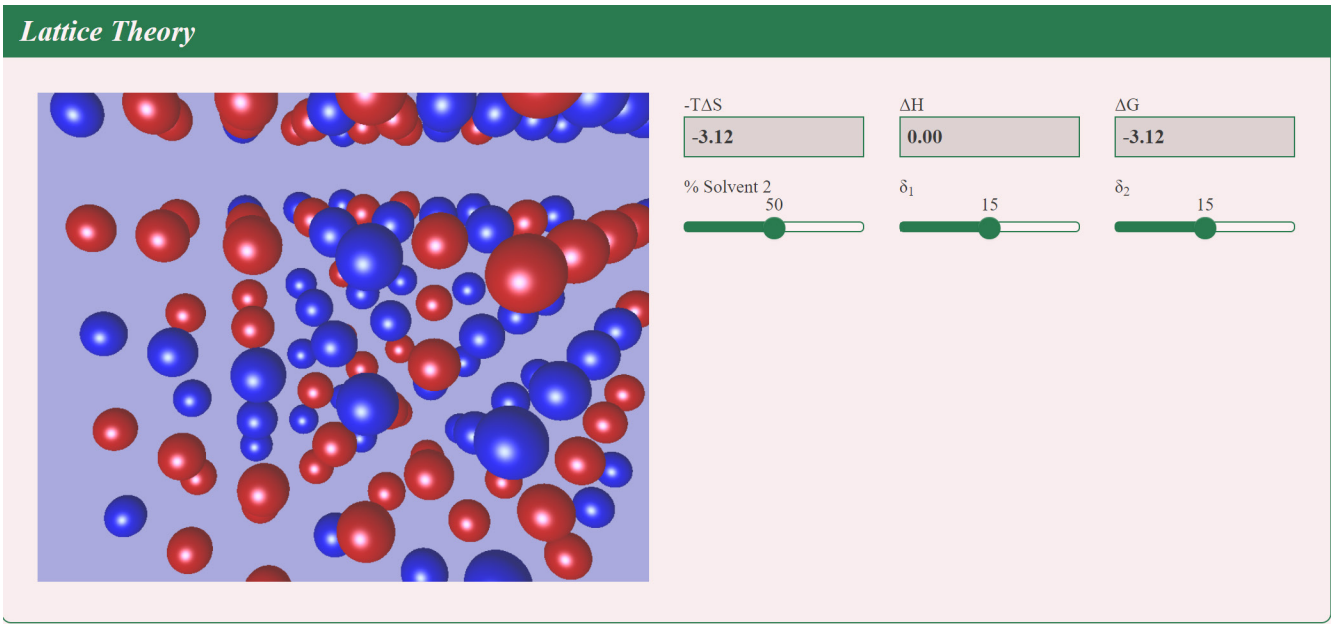


Figure 3-1 The lattice with a random 50:50 mix of molecules

In this example the red and blue molecules have been given the same properties (the δ values which will be explained shortly) so there is no enthalpic penalty from replacing lots of 1-1 interactions with lots of 1-2 and 2-2

interactions. Instead we have gained in free energy because we have more disorder (the random arrangement of blue and red molecules) and our change in entropy, TΔS, is -3.12, so our ΔG is also -3.12. This gain in (i.e. lowering of) free energy simply by creating disorder is the key driver of solvent/solute mixing, i.e. of solubility in general. Polymers are less easy to dissolve because the gain in entropy is far less because the monomer units within the polymer gain relatively little extra freedom.

This section is about Regular Solution theory and Lattice theory. All that the first term means is that we can calculate the entropic term simplistically (as if we were dealing with an ideal solution) whilst attending to a non-ideal enthalpic term. The Lattice theory part is just a way to let us calculate entropic and enthalpic terms easily.

So what about enthalpy? In this toy world, enthalpy can at best be neutral (which is what we've arranged here) and in general will make things worse. So our -3.12 gain in free energy is as good as it gets.

To understand the enthalpic effects let us start with something we know (in principle) about each molecule, which is how much energy it takes to remove all the molecules from the lattice and place them in the vapour state, the enthalpy of vapourisation. What we have done in the above example is to say that the enthalpies of vapourisation of the different molecules is the same. To remove 50% of the blue molecules and replace them with red molecules can only be enthalpically neutral if the energy lost by evaporating those blue molecules is recovered when we condense the red molecules. Another way to talk about this is via the *cohesive energies* of the molecules which are basically the enthalpies of vapourisation because these are the energies needed to break the cohesion in the lattice. It turns out that cohesive energies on their own are not too useful. Hexane and decane have very different cohesive energies because decane is a bigger molecule. Yet we know that in solvency terms they are very similar. Instead, let us talk of the *cohesive energy density* which is cohesive energy divided by MVol. Now hexane and decane are rather similar, as we would expect.

Finally we define the SP δ which is the square root of the cohesive energy density. The units are the rather inconvenient $MPa^{½}$. If our two molecules are very similar (say hexane and decane) then δ_1 will be similar to δ_2, if they are very different (say hexane and ethanol) then the δ values will be very different. And we can justify the act of "not caring about the enthalpy values" in the first example because whatever the magnitude of the δ values, as long as they are equal there is no enthalpic penalty for mixing.

Now we can introduce some equations into our lattice model. Remembering that the change of free energy, ΔG is made up of two terms, ΔG = ΔH-TΔS we

see, first, the entropic term which depends on the number fraction, n, of each molecule:

Equ. 3-1

$$\Delta S = n_1 \log(n_1) + n_2 \log(n_2)$$

Then we have the enthalpic term which uses the SP plus the number z which we can ignore, it is just the number of molecules around each molecule in the lattice, in this case 4:

Equ. 3-2

$$\Delta H = -n_1 n_2 z(\delta_1 \delta_2 - 0.5(\delta_1^2 + \delta_2^2))$$

When $\delta_1=\delta_2$ the ΔH term is zero and ΔG is made up only of the (helpful) entropic term. When the molecules are very different (hexane and ethanol) then the δ values (derived from the enthalpies of vapourisation) are very different so that the ΔH term becomes large and positive, which means that the free energy of mixing can become positive, which, of course, means that mixing (or solubility) is unfavourable. Taking the same 50:50 mix as before, but changing the SP parameters to 14 and 16.5 we find that the entropic advantages are quickly overwhelmed:

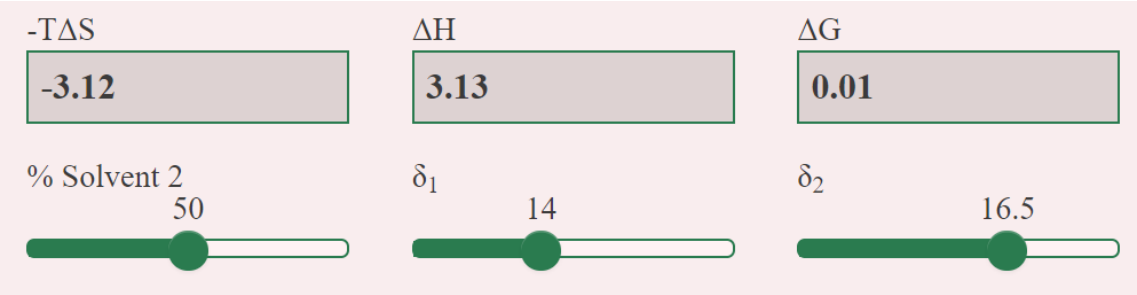


Figure 3-2 On the borderline of solubility. The different δ values give an enthalpic loss

What is happening is that the enthalpic gains from interactions between 1 and 2, $\delta_1\delta_2$ are overwhelmed by the loss of the 1-1 and 2-2 interactions in the average of δ_1^2 and δ_2^2.

A large part of solubility theory is taken up by finding ways to calculate the “bad” (or occasionally “good”) effects of 1-2 interactions compared to the loss of the “good” interactions of 1-1 and 2-2. Multiplying the two different terms to give $\delta_1\delta_2$ is the so-called “geometric mean” approximation compared to the normal average of δ_1^2 and δ_2^2.

And that’s it for the theory. We could spend a lot more time looking at subtle refinements, but it makes little difference. Why bother with subtleties when the whole model is a toy world? The take-home message is that if we know the

cohesive energy densities of molecules then with some rather simple accounting of enthalpic differences combined with some trivial entropic calculations, we can get a reasonable estimate of what happens when we try to mix or dissolve molecules.

The situation can be made to look slightly more respectable in terms of KB. One of the key equations we saw early on had (G_{11}+ G_{22}- $2G_{12}$). This is capturing the same issue - the balance of 1-1, 2-2 and 1-2. Making regular solution theory respectable requires a proper link between the rather naive assumptions of lattices etc. and the true picture as captured in Kirkwood-Buff Integrals. That is a work in progress[23].

Now we have the simple outline of a theory, we have to abandon the pretext that everything about a molecule can be captured in a single SP.

3.2 Hansen Solubility Parameters

Once the initial enthusiasm amongst formulators for the Hildebrand SP approach had died down upon the realisation that it just didn't work (and, indeed, nor could it have worked), the obvious thing to do was to remove the assumption that molecules could be described with a single parameter and add one, two or even three more. Adding just one meant that molecules had a general "dispersion" element (the basic van der Waals interactions) plus a "polar" element. This approach did not achieve much because something like ethanol has a polar parameter similar to something like acetonitrile even though they are very different solvents. Adding three parameters seemed an excellent idea, giving: dispersion, polar, H-bond acceptors and H-bond donors. Few chemists would want much more. Unfortunately, despite many efforts to make this highly-logical approach work, the donor/acceptor interactions proved too hard to implement. Not only did it require the determination of these parameters, formulae equivalent to the geometric term for cross-molecule donor/acceptor interactions had to be found. There has never been agreement on how best to do this just for two molecules and everything gets worse when trying to handle solvent blends.

So we end up with the slightly unsatisfactory compromise of adding just two terms to give a total of three parameters: dispersion, polar, and hydrogen bonding. On the positive side, this compromise has been wildly (unreasonably) successful for 50 years. Breaking the total cohesive energy δTot into three parameters poses no intellectual problems in terms of Regular Solution theory. One just makes sure that the sum of the squares of the three parameters equals δTot. The challenge lay in working out what those parameters would be for all relevant solvents, solutes, polymers and, as it turns out, pigments, nanoparticles etc.

If we had to do the task today we would start with a large dataset of properties such as enthalpies of vapourisation from which to extract δ_{tot} then would use

23 Seishi Shimizu, personal communication

some neural network that knew about the functional groups of each molecule. The neural net would get some self-consistent set of values which bootstrapped onto notions such as refractive index being strongly linked to the dispersive component, try to link the polar component with dipole moment and attempt some correlation with whatever groups can provide H-bonding including small contributions from aromatics as well as the obvious -OH groups. Indeed, in the HSPiP software described later, this sort of approach helps greatly in expanding the quality and quantity of HSP values. But Hansen had little computing power and there was no concept of "big data" so he more-or-less bootstrapped his way to a set of values using “all available means”. What is remarkable is that the majority of those values have stood the test of time.

Once the basic set was established it became possible to measure the HSP of polymers. Because polymers have no meaningful enthalpy of vapourisation there seemed no hope of establishing HSP values for them. But using the core idea which makes HSP so useful across so many areas, the HSP “distance”, it became routine to measure the values not only of polymers but of pigments, nanoparticles and crystalline solids.

So now we should find out about this core idea of distance, linking it to the earlier ideas of regular solution theory.

3.3 Distance

Let us play with the idea that we can define a Distance between two molecules based on their (for simplicity) Hildebrand SP. We define the distance, D, as a simple Cartesian distance $D^2 = (\delta_1 - \delta_2)^2$. The hypothesis is that this captures a key thermodynamic aspect of these two molecules and we can readily see what that is if we expand it: $D^2 = \delta_1^2 + \delta_2^2 - 2\delta_1\delta_2$. This is exactly the term used in the Regular Solution theory to calculate the enthalpic effect, balancing the gain from 1-2 interactions with the loss of 1-1 and 2-2 interactions. So when D is small, there is a very small enthalpic problem and when it is large the two molecules are unhappy in each other’s presence.

We can now properly introduce HSP and the Distance parameter with all that it entails. We have the three parameters δ_D, δ_P, δ_H representing Dispersion, Polar and H-bonding respectively. We know that, by definition, $\delta_{tot}^2 = \delta_D^2 + \delta_P^2 + \delta_H^2$ but that’s of no great use to us. Instead we are interested in the distance between two molecules which *should* be defined as:

$$D^2 = (\delta D_1 - \delta D_2)^2 + (\delta P_1 - \delta P_2)^2 + (\delta H_1 - \delta H_2)^2$$

Equ. 3-3

but *actually* is defined as:

Equ. 3-4

$$D^2 = 4(\delta D_1 - \delta D_2)^2 + (\delta P_1 - \delta P_2)^2 + (\delta H_1 - \delta H_2)^2$$

The first definition is the same Cartesian distance as in the simple version and breaks down into the same differences between the 1-1, 2-2 interactions and the 1-2 interactions. The additional factor of 4 in the second definition has never been properly justified, but the privileging of the dispersion term provides two benefits - the fits to real-world data are better and the plots which result from it are spherical rather than ellipsoidal. These pragmatic benefits are well worth the slight controversy that has dogged the factor of 4 for 50 years.

A novel insight into the origin of the factor of 4 was discussed by Hiroshi Yamamoto at the HSP50 Conference[24] and at the time of writing further developments are in progress.

Let us now use the Distance for a real-world benefit. Suppose we test a polymer (or pigment, nanoparticle ...) in a set of solvents chosen to have HSP covering a wide range of HSP space. We find that it is soluble in some and insoluble in others. Or, more usually, we find that it is “happy” (soluble, swollen, dispersed) in some and “unhappy” (insoluble, unswollen, sitting at the bottom of the tube) in others. From these simple experiments we can work out the HSP of the polymer. We can, for example, make a guess as to the HSP, then calculate the Distances to all the solvents. If our guess is right then the Distances to all the good solvents will be relatively small and those to the bad solvents will be relatively large. We can even define a Radius which defines the resulting sphere - all good solvents being inside that radius and all bad ones outside it.

Clearly we cannot get such a happy result via guessing, but it is not hard to find an algorithm to search HSP space for polymer values that give the best possible distinction between good and bad solvents. And this is how we measure the HSP of unknown solids. And, this, incidentally, is the reason that HSP, for all its faults and uncertainties, has proven so useful. Other theories might have intellectual advantages, but there isn’t one that makes it so easy to determine the key values of new materials, especially those (as is very much the case in real life) where the chemical nature of the material is unknown. Take, for example, a pigment particle with a dispersant. There is no obvious way to know what the properties of the “true” particle are, nor how the dispersant (even if we know its HSP) interacts with the particle and the environment. So the pragmatic measurement of the effective HSP of this complex is about as good as it gets. Hansen had some initial doubts about the validity of this approach which, after all, is stretching Regular Solution theory to its limits. But at the time he was working for a major paints company and there was no doubt about how useful

24 Hiroshi's talks on the "split δD" explanation for the factor of 4 as well as insights into donor/acceptor can be downloaded from https://www.hansen-solubility.com/conference/papers.php.

these measurements were in practice, so he promoted the technique in general and it has stood the test of time.

3.4 HSP Values

So far we have spoken of HSP in the abstract. Another advantage of HSP is that their values make intuitive sense. After a while, an HSP user can make a reasonable guess as to the values of a relatively straightforward molecule and not be too far wrong. We will look at the three parameters one value at a time.

- **δD** represents the dispersion or van der Waals component of a molecule and for simple molecules it represents the whole of HSP. We know that even a boring molecule such as hexane has a significant enthalpy of vapourisation and we know that δD must (via the square root of that enthalpy) be a modest, finite value. In fact for hexane that value is ~15MPa^{½}. For solvents that have very low enthalpies of vapourisation such as silicones or fluorocarbons, δD can drop to 11 or 12. Cyclohexane is more compact (higher density) and more self-associating so its δD is nearly 17. Aromatics happen to have some H-bonding component but they are mostly δD and we know that they self-associate (π-π*) and are harder to evaporate so it is no surprise that benzene's δD is 18. Once we start adding chlorine or sulfur atoms to an aromatic we can easily be up in the low 20s. And that's about it for δD, other than to point out that it has a strong correlation with refractive index which is, at a deeper level, a correlation with the polarizability of a molecule. Those molecules with more electrons able to move freely at the surface have a higher polarizability, a higher refractive index, stronger van der Waals interactions and a higher δD. δD is often dismissed as boring, yet it is often the dominant HSP component because *all* molecules have van der Waals attractions while polar and H-bond components, when present, can often be relatively small.
- **δP** is strongly related to dipole moment, so we know that hexane will have a value of 0 and that something with a large dipole moment such as acetonitrile will have a large δP. It turns out to be 18 and it then is no surprise that DMSO is ~16, acetone is 10, THF is 6 and so forth. You might not have got those exact values, but the general trends make sense and there are no big surprises. A symmetrical molecule such as 1,4-dioxane has no dipole moment. In this case the correlation breaks down because dioxane is not a "non-polar" molecule.
- **δH** is, once again, zero for hexane and for methanol it is 22. This allows us to scale ethanol and propanol as 19 and 17 because their -OH is "diluted" with CH_2 groups. DMSO is a good H-bond acceptor with a value of 10 and acetone has a value of 7. When Hansen did his original experiments he found that he had to give a small value of δH to aromatics like benzene and toluene. This "mistake" was the cause of some criticism but Hansen stuck with experimental data and kept those values. It was, therefore, with some satisfaction when, 40 years later, the small but significant H-bonding abilities

of aromatics were formally recognised by IUPAC in their definition of the H-bond. The values for simple aromatics are ~2.

Thus all three sets of values conform to a chemist's intuitions. There is a great benefit to this. There are many systems for estimating HSP from molecular structures and many of them do a reasonable job in their comfort zone. But try a molecule with a somewhat different structure and the system can provide ludicrous values. On the day of writing this I happened to use HSPiP's high-powered estimator of a very complex anti-oxidant molecule containing multiple -OH groups. The δP and δH values both came out as 0.1. Clearly this molecule had thrown the prediction algorithm out of its valid prediction zone. I also happened to read a paper on the HSP of biopolymers. One of them was given a δD of 8. It should have been obvious to authors and reviewers that this was impossible. These are extreme examples but it makes the point that because the HSP values conform to chemists' intuitions, there is a chance that many errors can be spotted in time. In each case, with common sense some reasonable values could be substituted for the illogical ones. The erroneous δD may be from some published value in old-fashioned $(cal/cm^3)^{½}$ which are a factor of ~2 smaller. A most egregious lack of common sense in another publication (I will not quote the reference) was to point out that one molecule failed to fit whatever trend was under discussion. The values used for that molecule were clearly the old units, and if they had merely multiplied them by 2 to bring them in line with the other units they would have found that the molecule fitted their trend.

3.5 Two more solubility theories

Via the HSP Distance we have a definition of "happiness" of a solute in a solvent. But we haven't said how that happiness impinges on actual solubility. Fortunately we have two rather robust theories which (for all their faults) give us a good idea of what to expect.

3.5.1 Ideal Solubility

The first theory is to do with crystalline solids. Let us suppose that we have a perfect solvent in terms of lattice theory, so that the solvent-solvent and solute-solute interactions are equal to the solute-solvent interactions. Unless we have extra effects (such as strong donor/acceptor interactions) we cannot get a better solvent. So, is the solute infinitely soluble in the solvent? Our experience tells us that there are plenty of solutes that are, effectively, brick dust which no solvent can touch and that there are others that seem remarkably soluble in a wide range of solvents. The key to this is the idea of Ideal Solubility - the solubility of a solute in an ideal solvent - defined in the terms above where 1-1 and 2-2 interactions equal 1-2 interactions, in other words with an activity coefficient of 1 at all concentrations. The word "ideal" does not mean "the solvent best able to get this solute into solution". There are situations where a solvent which has a much stronger interaction with the solute (e.g. a basic solvent with an acidic

molecule) has an activity coefficient of <1 but in many ways that is changing the solubility issue from one molecule to another. There are borderline cases where there are strong donor/acceptor effects between solute and solvent. They definitely help but are not the focus of this section. Our definition of "ideal" is "activity coefficient = 1".

The key to understanding the issue of ideal solubility is that lattice theory is for compatibility of *liquids*. So to dissolve our crystalline solute we first have to make it liquid - i.e. we have to virtually melt it. We know intuitively that high MPt solids will be harder to melt virtually and we can also imagine that solids with a high enthalpy of fusion ΔH_F will be harder to melt. Those with advanced understanding will recognise that changes of heat capacity between the real solid and the virtual liquid will also be important. The thermodynamics capture these intuitions precisely in the Ideal Solubility equation which gives us the mole fraction, x, of the solute in the ideal solvent at temperature T when its MPt is T_m, and it has the enthalpy of fusion ΔH_F and change in heat capacity of ΔC_p :

$$R\ln(x) = \Delta H_F(\frac{1}{T_m} - \frac{1}{T}) + \Delta C_p(\frac{T_m}{T} - \ln(\frac{T_m}{T}) - 1)$$

Equ. 3-5

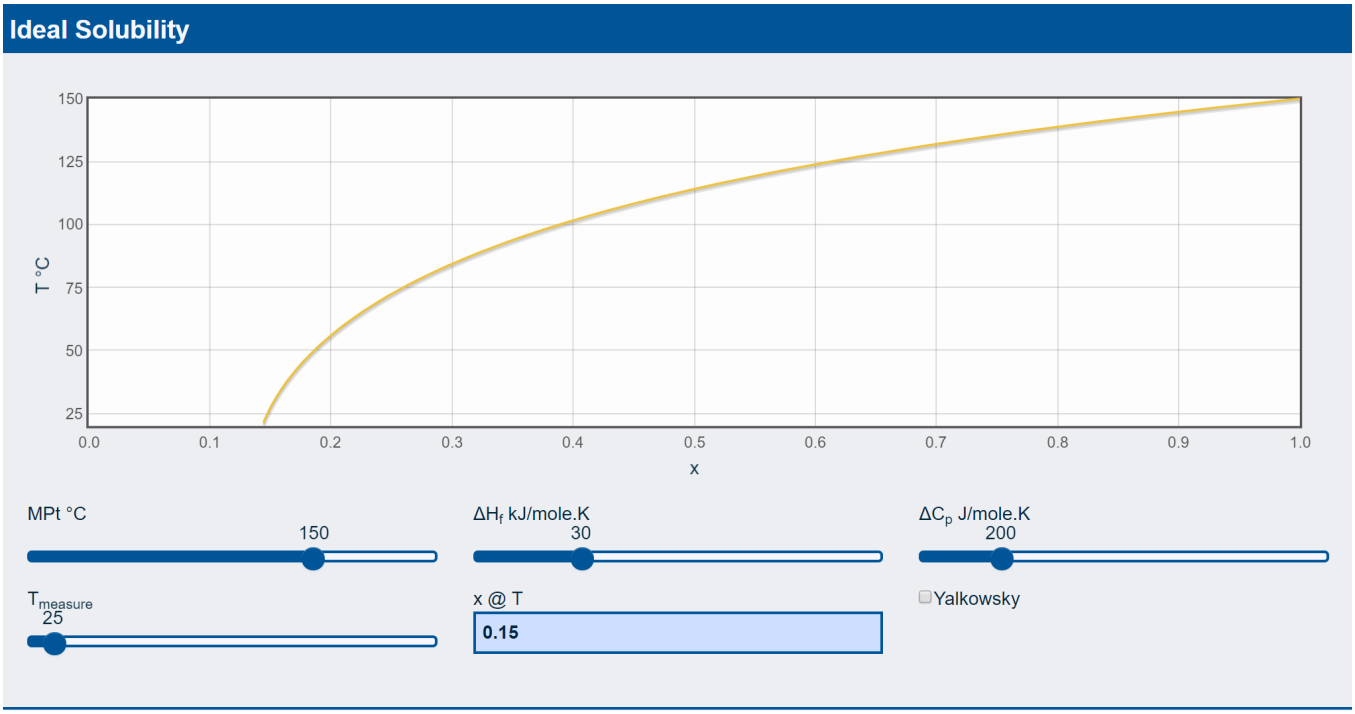
Although this equation is correct, it is generally unusable as we are unlikely to know the enthalpy of fusion and the change in heat capacities. Fortunately, the careful analysis of this unhappy situation by Yalkowsky[25] tells us that the uncertainties in our knowledge of the two uncertain parameters is generally large enough that we can forget about them and use the delightfully simple alternative for which we only need to know T_m:

$$\ln(x) = -0.023(T_m - T)$$

Equ. 3-6

These ideas can be explored in the Ideal Solubility app:

25 Samuel H. Yalkowsky and Min Wu, *Estimation of the Ideal Solubility (Crystal–Liquid Fugacity Ratio) of Organic Compounds*, Journal Of Pharmaceutical Sciences, 99, ,1100-1106, 2010



App 3-2 https://www.stevenabbott.co.uk/practical-solubility/ideal.php

The graph tells us what the mole fraction solubility is in an ideal solvent, at any temperature up to the MPt where, by definition, the solute is perfectly soluble as they are miscible liquids. The app allows you to specify all three parameters, though the Yalkowsky option is highly recommended if you only know the MPt.

The point is not so much the calculations but the realisation that before you even bother with HSP (or, as discussed in the next chapter, COSMO-RS) you should have an idea of what the ideal solubility will be. If you absolutely require x=0.2 for your application and the ideal solubility is 0.1 then HSP cannot help you. If you require x=0.05 and the ideal solubility is 0.3 then you don't have to be too fussy about what solvent to choose. Or, if you have the 0.2 requirement with the 0.1 ideal solubility, then you need to find some positive (non-ideal) way to encourage the solute into the solvent via clever donor/acceptor or acid/base interactions.

3.5.2 Flory-Huggins

The second theory is necessary for understanding polymer solubility. Here we assume that the polymer is 100% non-crystalline. As soon as you add polymer crystallinity then the equivalent of ideal solubility has to be factored in and life gets too complex.

As we saw for non-polymers, solubility is a mix of entropic and enthalpic effects. The problem for polymers is that the enthalpic effects are similar to those of small molecules, i.e. they seldom help, are at best neutral and usually hinder, while the entropic effects are much smaller. All the monomers linked together through a lattice are guaranteed to have less entropy than their equivalents

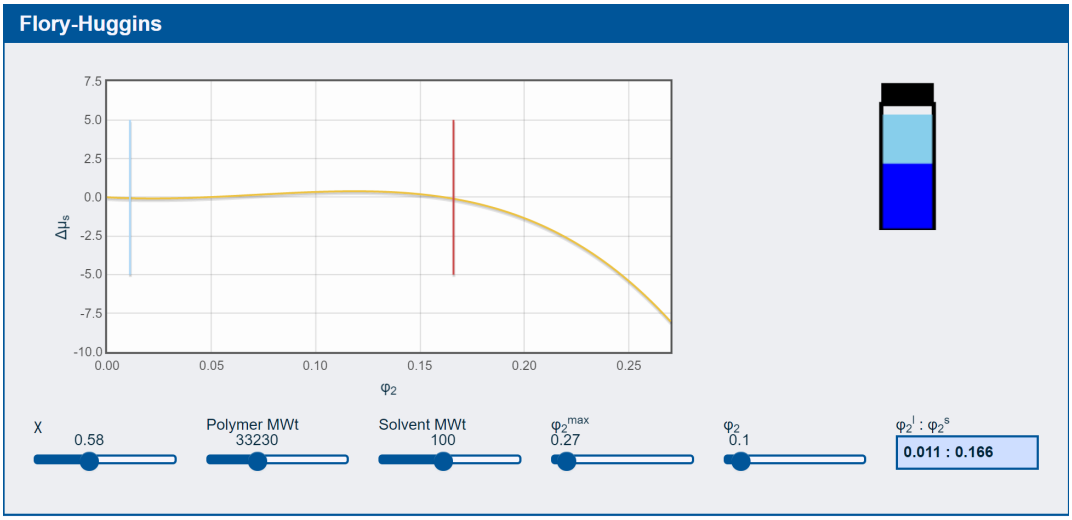
sitting as individual monomers on the lattice, and the gain of entropy by mixing with the solvent is rather small.

This means that the solubility of *polymers in solvents* is often not very large. Although this is true, we can be confused by another fact which is that the solubility of *solvents in polymers* can often be large. Such facts are not often put that way. Instead, polymer physicists talk about "spinodals" which, while being correct, is less useful to most of us than knowing that dissolution of polymers (*polymers in solvents*) is much harder than swelling (*solvents in polymers*). The key equation is called Flory-Huggins and is expressed in terms of volume fraction φ because mole fraction makes less sense for polymers. Where 1 is the solvent and 2 is the polymer and the enthalpic interactions between polymer and solvent can be defined by the χ (chi) parameter (to be discussed shortly), we find that the change in chemical potential (a negative change is good for solubility) is given by:

$$\Delta\mu_s = RT[\ln(\varphi_1) + \varphi_2(1-\frac{1}{x}) + \chi\varphi_2{}^2]$$

Equ. 3-7

The factor x is the ratio $MWt_{polymer}$/ $MWt_{solvent}$. When this is large (large MWt polymer) 1/x tends to 0, meaning an increase (bad) in the φ_2 contribution to the overall the entropic contribution. Clearly, the larger χ, the more positive the effect of φ_2 and the worse the solubility. What is not clear is that the shape of the curve of $\Delta\mu_s$ can have a subtle dependence on φ_2. For low χ values (below the magic number of 0.5) the polymer is soluble at all volume fractions. For high χ values (say >0.7), Δμ increases so rapidly that it is obvious that the polymer is not soluble. At intermediate χ values the curve dips negative then goes slightly positive then negative, as in the screen shot from the Flory-Huggins app:



App 3-3 https://www.stevenabbott.co.uk/practical-solubility/polymer-solubility.php

At a critical inflection point, shown by the first vertical blue line, the polymer is no longer soluble in the solvent. At a second critical inflection point, marked with the

red vertical line, the polymer and solvent are again compatible. If we have (as shown in the input) a φ_2 of 0.1, which is between these points, what do we find? The two points are the "spinodal" points which means that the system splits into two. Two phases appear in the test tube, with the dilute polymer solution (light blue) on top and a swollen polymer on the bottom. Here's the important bit. The *concentrations* of the two phases are fixed by the shape of the Flory-Huggins curve. In this example there is 0.011 of polymer in solution and 0.188 in the swollen phase, i.e. the polymer contains 0.812 solvent. What happens if we have 0.15 of polymer? We still have two phases in the test tube and they are still 0.011 and 0.188, but we have a different *ratio* of the two - more of the lower phase and less of the upper. Try it in the app to get a feel for this. The spinodal splits systems into two phases, their *concentrations* are fixed for any starting concentration between those two points, only the *ratio* of the two phases changes. For those who think that the splitting points should be the minimum and maximum of the curves, you are partly right - these are the "binodal" points but in practice the spinodals win.

There are two important practical points here. First, although this polymer is scarcely soluble in this solvent (you can only get 0.011 volume fraction), you can still get 0.812 solvent into the solvent. It is very easy to make the assumption that if the polymer is rather insoluble, the solvent won't "touch" the polymer. Second, when you play with the app you find that small changes of χ or of MWt (which affects 1/x) can flip the system from soluble at all fractions to rather insoluble (though with lots of swelling). A formulation which seems to be fine can flip to unworkable if you happen to have been in a zone with χ close to the critical value and if there is a small change to the solvent (blend) or a change in supplier of the polymer which means a somewhat higher MWt.

There is a further subtle point to do with language. We have a strong intuition that we know what "solubility" means. So when we have a polymer and solvent system where the polymer is not completely soluble we say that it is "insoluble" and think we know what this means. The spinodal example shows that we have one phase which is nearly 20% polymer. By no normal definition of "insoluble" can you have a 20:80 mix of polymer:solvent. The solvent and polymer, therefore, are not desperately unhappy in each other's presence on this side of the spinodal. The final chapter discusses why our (mis)use of solubility language so often leads us astray, and makes positive proposals of how we can avoid falling into these linguistic/scientific traps.

3.5.3 χ in practice

Clearly it would be a good idea to know χ for our polymer/solvent system. If you happen to have a lot of time and money, plus a neutron source, you might be able to measure χ. But as a splendid review ("Beware the Flory parameter")

shows[26], the majority of such χ values are, to put it politely, of dubious value. Instead, let us use a straightforward way to work out χ in any system that interests us. Hansen showed a long time ago that we can replace χ with the HSP Distance, D, via:

Equ. 3-8

$$\chi = \frac{MVol.D^2}{4RT}$$

Given that it is trivial to calculate D for any solvent (the MVol term is that of the solvent) if the HSP of the polymer is known, we can very readily work out where we are in terms of Flory-Huggins. This raises the question of why the difficulties raised in the Miquelard-Garnier review can be so readily overcome by a relatively simple theory. The answer is that the act of measuring the HSP of the polymer is defining the border between soluble and insoluble (or, perhaps, between single phase and spinodal), so the uncertainties about the polymer, its purity, its MWt distribution etc. are all swept up into the measurement itself.

This means, as HSP users regularly find, that you cannot just take "the" HSP for a polymer and use it. It does not require much of a difference in % head-to-head polymer or "helpful" co-polymer or slight degree of cross-linking etc. to make one sample of polyX behave significantly differently to another sample - at least near the critical spinodal zone. A very good solvent for one polyX will be a very good solvent for the other polyX. And the same, in reverse, for a very bad solvent. But a "just about good enough" polymer for one polyX might be rather useless for another and a "borderline bad solvent" for one version might be borderline good for another version. As the Flory-Huggins app shows, near the spinodal the chemical potential curve is balanced delicately so that a small effect in either direction can tip it one way or another.

You are always welcome to be a purist and spend a year studying the neutron scattering of a polymer in a few (deuterated) solvents. But a few days work with a bunch of solvents to calculate the HSP sphere and you have much of the information you need for practical formulations.

3.5.4 χ and KB

Although this chapter is about lattice theory, it is interesting to see what happens when χ>0.5 using the KB approach. The broad-brush entropy/enthalpy approach from Flory-Huggins gives no insights into what the molecules are doing before and after the spinodal separation. The reason KB is used throughout the book is that it gives an intuitive picture of what is happening. So let us see what that picture is in this case.

---

26 Guillaume Miquelard-Garnier and Sébastien Roland, *Beware of the Flory parameter to characterize polymer-polymer interactions: A critical reexamination of the experimental literature*, European Polymer Journal 84 (2016) 111–124

Following the logic of a paper by Horta[27], we start with something that can be observed in principle, the zero angle scattering structure constant which, for simplicity, I will call S though it is really $S_{cc}(0)$. The "cc" is "concentration-concentration" and the fact that our KBI values are "concentration-concentration" fluctuations is not a coincidence. As discussed in the KB chapter, the relationship between scattering and KBI is a precise one and those skilled in the art can rapidly change between them when necessary. In our case, Horta points out that we can get S from χ and from S we get the G_{ij} values via the following sequence, where the MVols of the solvent, 1, and polymer, 2, are V_1 and V_2 and where $r=V_2/V_1$:

Equ. 3-9

$$\frac{1}{S}=\frac{1}{\varphi_1}+\frac{1}{r\varphi_2}-2\chi$$

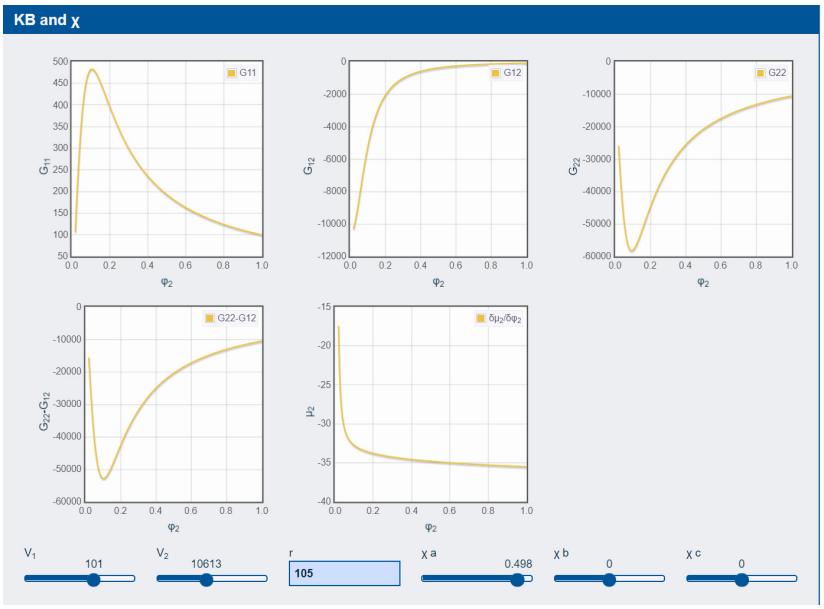
Equ. 3-10

$$G_{11}=\frac{V_1}{\varphi_1}\left(\frac{S}{\varphi_1}-1\right)\quad G_{12}=-S\frac{V_2}{r\varphi_1\varphi_2}\quad G_{22}=\frac{V_2}{\varphi 2}\left(\frac{S}{r\varphi_2}-1\right)$$

Equ. 3-11

$$\frac{\delta\mu_2}{\delta c_2}=\frac{1}{c_2\left(1+c_2\left(G_{22}-G_{12}\right)\right)}$$

The screen shot shows how this works within an app



App 3-4 https://www.stevenabbott.co.uk/practical-solubility/kbchi.php

From the change of chemical potential with concentration a little bit of arithmetic gives us the dependence of μ on φ_2 which is basically what we plotted in the

27 Arturo Horta, *Nonrandom Distribution of Molecules in Polymer Systems. 1. Theory and Model Calculation, Macromolecules* 1992, 25, 5651-5654

conventional Flory-Huggins app. In the screen-shot, the system is just poised to flip into the spinodal state because χ is just below 0.5.

Now we get a chance to see what is happening. The solvent is bunched up on itself, with the large positive G_{11}. The solvent and polymer are unhappy together, shown in the large negative G_{12}. The most significant behaviour is the polymer's. At a smaller χ it plunges to a large negative value, thanks to the excluded volume of such a large molecule. Yet now it is turning positive because it wants so little to be associated with the solvent. Looking at the G_{22}-G_{12} term in the derivative equation, as soon as G_{22} becomes larger than G_{12} which itself is negative, the derivative becomes positive and the system starts to flip. You have to play with the app to see all this in detail. At the point of flipping, KB theory breaks down and the curves become rather too wild to be informative. But we have already achieved something significant. Just with two MVols and a χ we can see at the molecular scale what is happening to cause a system to flip. Although the flip is sudden, it does not just "happen" as Flory-Huggins suggests. We can see that the solvent and polymer are already starting to self-associate. The flip is just the result of self-association going a bit too far.

There is one other point. The G_{ij} values in the range of φ_2 = 0.3-1 show no special features. The polymer and solvent have no problems co-existing in this range. Although by definition they don't much like each other, even at 50:50 G_{22} is more negative than G_{12} so the system is stable. Why is the polymer-polymer self-association *worse* than polymer-solvent? It is simply the excluded volume effect. It is *much* easier to have a bunch of solvent molecules close to the polymer chain than another polymer chain. So the solvent is happily soluble in the polymer in this higher φ_2 range. This sounds strange, because we "know" that at high χ values the polymer is insoluble. This takes us back to the earlier discussion about solubility language. The reason we don't habitually think that the polymer is happily soluble in the solvent is that in this domain we only talk about the solvent "swelling" the polymer. By using a different word we frame our thoughts away from the word "soluble" so we don't even question whether we can say that the solvent is soluble in the polymer. And in any case, we cannot do conventional solubility experiments because the viscosity of the system is too high, so we just do swelling experiments as if they are some totally different phenomenon.

While on the subject of language, when it comes to water and somewhat hydrophobic polymers the language tends to imply that something dramatic is happening when the water is sitting next to the polymer. Although the topic will be discussed at length in the Language chapter, we can already see how the previous paragraph helps explains why water effects might not be as mysterious as they seem. The polymer, because of excluded volume, does not like being next to itself so just about anything will work just fine. And *all* molecules have mutual interactions via van der Waals forces so there is no problem for the water to be next to the polymer and therefore no mysterious "hydrophobic hydration"

force needed to explain it. Discussions then go on about the fact that the water next to the polymer has a different structure from bulk water. But *any* solvent in contact with a polymer is going to have a different structure. Yes, the difference is generally bigger when water is involved; but it is not some amazing new phenomenon. In general, the G_{22} excluded volume KBI is far larger than G_{11} or G_{12} and whether the balance of KBI tips from negative to positive is often a matter of subtle details that do not require explanations via vaguely-defined concepts such as "hydrophobic hydration".

3.6 Polymer-Polymer (in)solubility

If you attempt the Flory-Huggins formula with two polymers, the conclusion is rather shocking. It shows that for two polymers of degree of polymerisation of 1000 (~50K MWt), the χ parameter above which they are immiscible is 0.002 and therefore the HSP Distance for miscibility is <0.1. The precise figures don't matter because the conclusion is inescapable: most polymers are immiscible. When I first heard this I thought it was just wrong. Surely you can take PMMA and PEMA and make a stable polymer blend. Their HSP values are respectively [18.6,10.5, 5.1] and [17.6,9.7,4] (Distance=2.4) which means that there are many solvents in which they are mutually soluble, i.e. they behave rather similarly, just as intuition would predict. This is true, but a high MWt blend of the two polymers phase separates if you heat it for a while - they really are immiscible. Indeed, it is worse than that. Even high MWt PE and deutero-PE are also immiscible because deuteration is enough to create a small enthalpic difference which the very small entropic mixing terms cannot overcome.

In practice one can readily make blends of PMMA and PEMA that are "good enough", especially if their MWts are not too high. The kinetics of phase separation are very slow. But if you heat such a blend for a prolonged period, they *will* phase separate.

And yet there are many rather stable polymer blends of polymers with very different HSP values. How can this be? The answer comes from Coleman and Painter's huge body of work on donor/acceptor polymer pairs[28]. As they point out, basic HSP can *never* have a χ value < 0. The fact that donor/acceptor polymer pairs are miscible can *only* be because χ < 0. And this happens if one polymer, for example polyvinylphenol, has the ability to donate a hydrogen and the other polymer, such as polyvinylacetate is able to accept it. So polyvinylphenol and polyvinylacetate are nicely miscible. The idea of donor/acceptor is intuitively clear though with some thought an objection arises. Polyvinylphenol form donor/acceptor bonds with itself - so why doesn't this win out compared to donor/acceptor with the acetate? Coleman and Painter did all the hard work of infra-red analysis of these systems to show that, indeed, the self donor/acceptor effects do diminish the interactions with the acetate, and that for any given donor/acceptor polymer pair there might be regions of

28 Michael M. Coleman and Paul C. Painter, *Hydrogen Bonded Polymer Pairs*, Prog. Polym. Sci., 20, 1-59, 1995

miscibility and immiscibility. If you have an interest in donor/acceptor polymer pairs then you need to study the Coleman and Painter approach in detail. The take home message for the rest of us is that HSP have a clear limitation - they cannot handle donor/acceptor effects. This has been known from the beginning (the heat evolved from mixing chloroform and acetone has often been cited as evidence against HSP). In the introduction to this chapter it was pointed out that a 4-D HSP would have been intellectually superior. But the gains (being able to predict chloroform/acetone or polyvinylphenol/polyvinylacetate) are modest compared to the difficulties faced by all attempts (including those of myself with Hansen and Yamamoto) to make a general-purpose 4-D HSP. The work on MOSCED[29] (Modified Separation of Cohesive Energy Density) which adds a donor/acceptor term showed great promise but despite heroic efforts from the Lazzaroni thesis[30] never created a sufficiently robust and usable methodology for anything other than some pure solutes in single solvents. The key issue is that there is no coherent explanation of what happens to the relevant donor/acceptor elements within and between solute and solvent. Chloroform/acetone is very simple because the first is a pure donor and the second is a pure acceptor. Coleman and Painter could do the hard work to see the balance of internal/external effects in polyvinylphenol, but such work was a concerted effort on some relatively well-defined systems. As we shall shortly see, HSP are immensely powerful in terms of solvent blends and I am not aware of any practicable system that can calculate the donor/acceptor complications for a two-solvent plus polymer system. For general-purpose solubility issues the simpler 3-parameter system works remarkably well *provided the formulator is alert to issues such as donor/acceptor effects specific to their own system.*

3.7 One more polymer-polymer formula

The fact that most polymers are immiscible is sometimes used as an excuse to deny the existence of one form of strong adhesion between polymers across an interface[31]. It is obvious theoretically that there will be a strong increase in adhesion (compared to mere surface energy) if polymer chains from one polymer can intermingle or entangle with those of the polymer on the other side of the interface. But those who believe (wrongly) that surface energy is important for strong polymer/polymer adhesion point out that most polymers are immiscible and "therefore" such intermingling and entanglement cannot take place. Fortunately this "disproof" of diffusion across the interface is, itself, easily dismissed. Although we all agree that polymers are not generally fully miscible, for adhesion we are not asking for miscibility - we are asking for there to be some mutual inter-penetration of polymer chains across an interface and

---

29 Eugene R. Thomas and Charles A. Eckert, *Prediction of Limiting Activity Coefficients by a Modified Separation of Cohesive Energy Density Model and UNIFAC*, Ind. Eng. Chem. Process Des. Dev., 23, 194-209, 1984

30 MJ Lazzaroni*, Optimizing solvent selection for separation and reaction,* PhD Thesis, Georgia Inst. Tech,, 2004

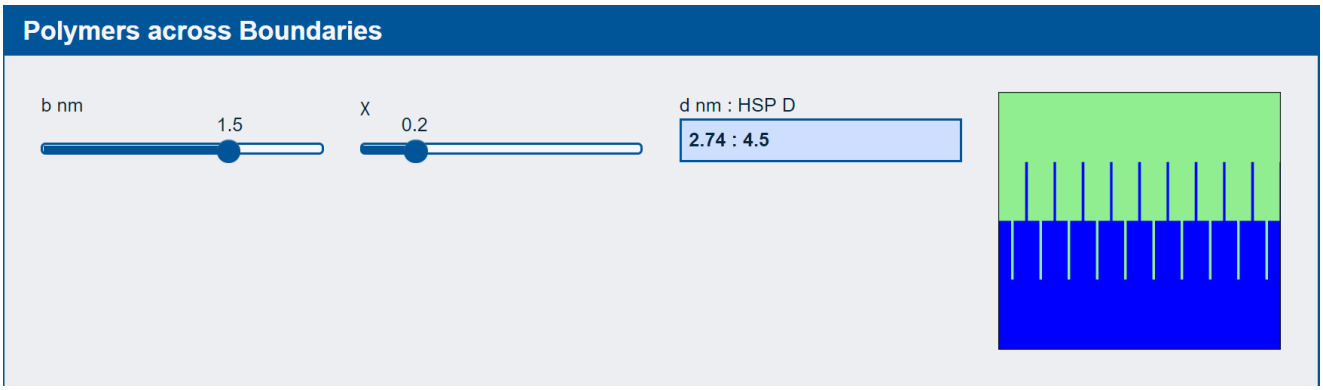
31 A full discussion on the uselessness of surface energy and the usefulness of intermingling/entanglement can be found in my Adhesion Science: Principles and Practice book and on my Practical-Adhesion website.

the thermodynamics of this are clear. The Helfand formula[32] (note that the title of the paper is all about interfaces between *immiscible* polymers) tells us that the distance, d, polymers of "effective" segmental length b (generally a few monomer lengths) can intermingle depends on the χ parameter or the HSP Distance, D:

Equ. 3-12

$$d = \sqrt{\frac{b}{6\chi}} = \sqrt{\frac{4RT.b}{600.D^2}}$$

The exact conversion of χ to HSP Distance isn't entirely certain (the 100 is a nominal MVol for a monomer unit), but the message is clear that a smaller HSP Distance gives a larger intermingling across the interface and, therefore, stronger adhesion between polymer phases. This applies to classic adhesion but also to compatibility of polymer blends.



App 3-5 https://www.stevenabbott.co.uk/practical-solubility/polymers-across-boundaries.php

The Miquelard χ paper referred to earlier has a good discussion on the applicability of Helfand. Although the theory is not perfect, the evidence is that it is remarkably good for practical purposes. It is strange that adhesion science courses will spend hours discussing relatively useless surface energy effects and do not even spend minutes discussing the implications of the Helfand formula and the practical power of HSP in understanding adhesion effects.

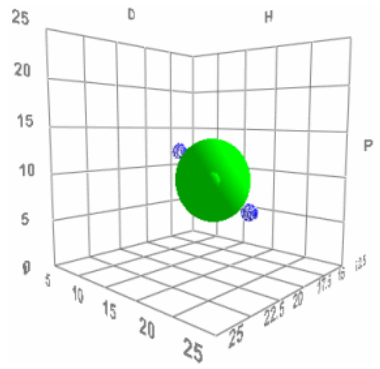
3.8 Solvent blends

If you have the HSP of a polymer, pigment, nanoparticle etc. and wish to find a good solvent for it, the process is simple: make a list of the HSP of all relevant solvents, calculate their Distances from the solute and whichever is the closest is the best solvent. Unfortunately, although "best" is correct in terms of solubility, it is rarely "best" in terms of cost, flammability, odour, safety profile or any of the other parameters of importance for a given formulation. It might be possible that by doing down the list the best balance of solubility versus the other properties

32 E Helfand, Y Tagami, *Theory of the interface between immiscible polymers*, Journal of Polymer Science Part C: Polymer Letters, 9, 741-746, 1971

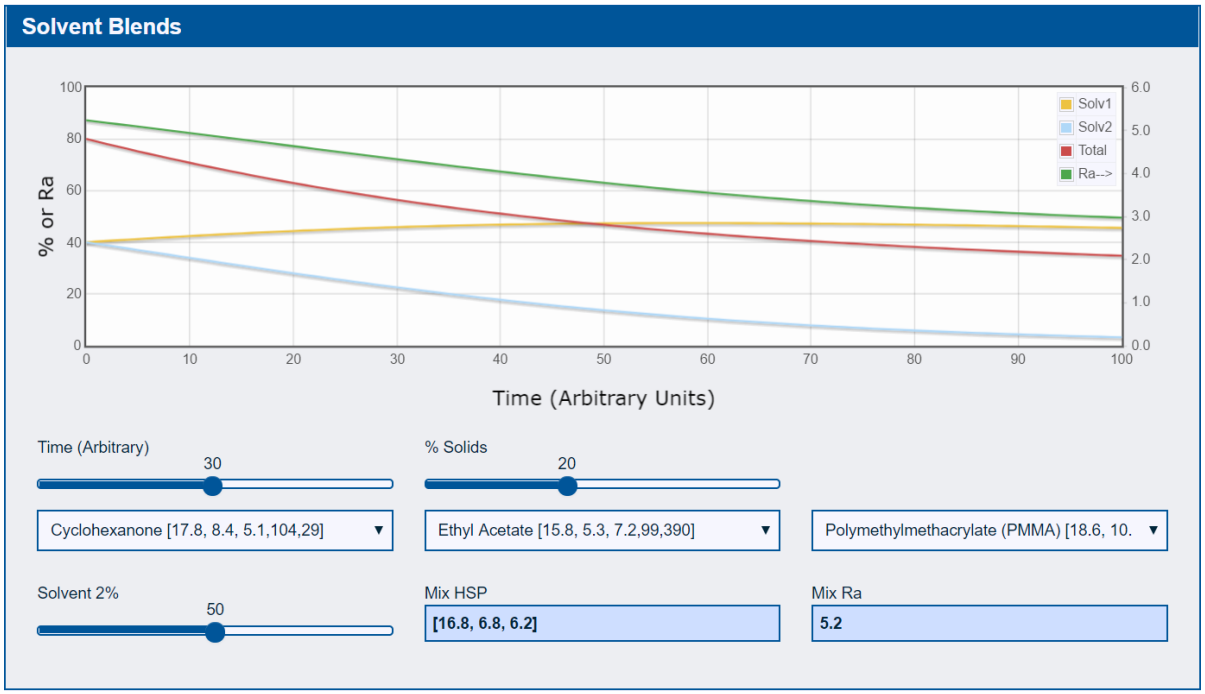
can be found and your search for the best solvent is over. It can equally happen that shortly after this choice has been made, a safety committee, or some consumers, or some government decides that that specific solvent is no longer desirable and you are left without a satisfactory system.

It can also be the case that a property such as volatility might be correct for the first part of a process but bad for the second part, or vice versa. So yet again, a single solvent simply cannot do what is needed.



Returning to the early days of HSP, it occurred to Hansen that one of the rules of HSP made a non-intuitive prediction. The rule is that the HSP of a mixture of solvents is the volume-averaged HSP of the solvents in the mixture. Hansen then imagined the situation shown in the diagram: a polymer with an HSP shown as the small dot in the middle, a radius of solubility shown as the bigger sphere, and two bad solvents outside the sphere, on opposite sides. It is obvious from the diagram that a 50:50 blend of these two bad solvents would have an HSP right in the middle of the sphere - creating a perfect solvent. This seemed scientifically sensible but intuitively false. Hansen realised that this was the ultimate test of his ideas so he looked through his sets of solvents and polymers and found an example with a bonus feature. The two (bad) solvents happened to have the same Hildebrand SP so according to Hildebrand theory, mixing them could have no effect on their inability to dissolve the polymer. Of course, Hansen's experiment worked out well - the mix of the two bad solvents produced a good solvent. He went on to find, and publish, many more such examples and a recent experimental demonstration of the effect is discussed below.

This principle of mixing solvents (good, bad or medium) to create an optimum blend is immensely liberating. For example, if there is a cheap, safe solvent that is not good enough for the solute of interest, adding the right amount of a more expensive solvent to bring it into the solubility zone is straightforward. Those who formulate coatings can do two different tricks. In some cases it is desirable to have the formulation "crash out" very quickly during evaporation. By choosing an excellent solvent with high volatility, very quickly after coating the mix changes to an unfavourable one and the solute falls out of solution. More frequently, the desire is to keep the solute as "happy" for as long as possible, so that the poorer solvent must have a higher volatility. This keeps the system mobile, allowing coatings to dry out with low stress (and therefore low curl) and with high gloss. If, as is often the case, the formulation contains multiple components with different HSP, the blend of solvents can be optimised to keep one or the other in solution for as long as possible. The example shows that as the more volatile, higher distance (shown as Ra), ethyl acetate evaporates from the initial 50:50 mix, the cyclohexanone is a better match (Ra ~3)



App 3-6 https://www.stevenabbott.co.uk/practical-solubility/solvent-blends.php

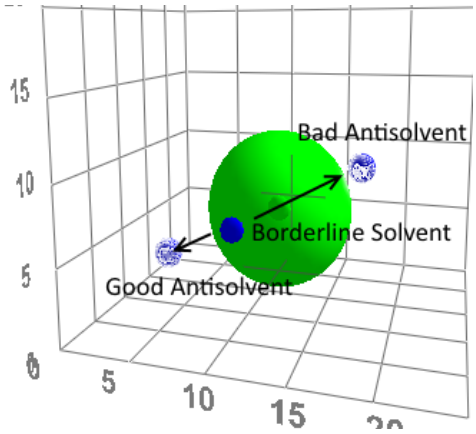
If, instead of wanting to match a solute, we want to match the HSP of a solvent, in order to replace it (e.g. replace dichloromethane) then the same principle applies - find the best blend of price/safety/volatility that matches the solvent you wish to replace. I once had to replace, for a cleaning application, one solvent blend (effective, volatile so easy to "dry", but unpleasant and unsafe) with another solvent blend. Unfortunately the only safe pair of solvents I could find were of low volatility, so cleaning/drying in a manner similar to the original blend was not possible. They were also larger molecules so were much slower to dissolve the difficult polymer. However, both these problems turned into an advantage. The whole, complicated, system that needed to be cleaned could be smothered with the non-volatile blend and we could attend to other tasks. Coming back to the system after some time, with the non-volatile solvents still there, it was easy to wipe everything clean then rinse with a very small amount of the original blend to allow the system to dry out.

So rational solvent blends are a key solubility tool that enable many new functionalities. As mentioned previously, the 3D-HSP might not be as perfect as a 4-D system with donor/acceptor, but the absence of a rational theory of 4-D solvent blends is a serious impediment for practical use.

There is a key problem in designing such blends. HSP has no way of knowing if two solvents are miscible or not, so HSP might suggest a wonderful solvent blend which is unusable in practice. This sounds like a severe limitation of HSP. In fact it is a general problem of solubility theory. Many readers will know of Wilson parameters for calculating activity coefficients. Famously, the Wilson approach "predicts" that all solvents, no matter how high the activity coefficients, are miscible at all proportions. Immiscibility is a very challenging issue and there

is also a severe shortage of high-quality datasets of (im)miscibility on which to base any prediction tool. Even the COSMO-RS solubility approach cannot be relied on to predict immiscibility[33] in "borderline" regions. This is because immiscibility is all about spinodals which, in turn, depend on subtle enthalpy/entropy balances. The word "subtle" means "in the 1kJ/mol range". There is no generally usable theory on the planet that can reliably calculate to this accuracy. The "failure" of COSMO-RS to predict immiscibility in borderline cases is actually a tendency to over-predict immiscibility in systems that just happen to be 0.5kJ/mol the other side of the immiscibility line. Examples to do with water miscibility are discussed in later chapters. Of course, COSMO-RS and HSP do a good job for pairs that are obviously miscible or immiscible as these are very far from the zones where 1kJ/mole makes any difference.

If even COSMO-RS cannot reliably predict immiscibility, those who use HSP to design interesting solvent blends have to accept that some of their prospective formulations will fail through a lack of miscibility.



The idea that the addition of a bad solvent can make a better solvent has an important negative consequence for those who wish to use anti-solvents to precipitate a solute at a chosen moment, e.g. for microencapsulation. If, as in the diagram, we suppose that we have a borderline good solvent then we can precipitate the solute with either of the two bad solvents. The "good" anti-solvent would be excellent because just a small amount will take the system outside the sphere. The "bad" anti-solvent would be a disaster because by adding it, the solubility will actually increase (heading to the centre of the sphere) before decreasing and finally precipitating the solute.

3.9 HSP and Temperature

The diagram above also helps to explain a curious phenomenon seen from time to time with polymers. Suppose you have a good, but borderline, solvent for your polymer. Now increase the temperature of the system. What happens? We all know that the polymer will be more soluble. Yet sometimes the polymer comes out of solution at higher temperature. When this happens in water for a polymer such as PNIPAM (discussed at length later), all sorts of special explanations are offered. Here we are talking about normal humble solvents where there are no known special effects. How does HSP help us to understand what is going on?

---

33 It is characteristic of the COSMO-RS team that they regularly point out the limitations of what a computer can do and provide users with detailed computational outputs to allow sanity checking of the results. They equally point out that experimental data contains flaws and that sometimes failures of the calculated results to match experimental data are clearly because the data itself is flawed. Computer tools are optimal when paired with a discerning human brain.

The key is that HSP are related to cohesive energy density. If you increase the temperature then the density goes down and, via well-known formulae, the three HSP all decrease. For a solvent, the density change is relatively large so the HSP changes are relatively large. For a polymer the change is much smaller and can be assumed to be zero. Using the diagram from the anti-solvent discussion we can now see what happens. The HSP of the borderline solvent are relatively low. So by raising the temperature the values will head towards those marked as the good anti-solvent, while the HSP of the polymer remains constant. So the borderline solvent will become a bad solvent. Had the borderline solvent been in the region close to the bad anti-solvent then raising the temperature would shift the HSP closer to the centre, making a borderline solvent a good solvent.

In general, increasing the temperature of a polymer system increases solubility because of the general temperature effect combined with the fact that most solvents will not happen to be in the critical low HSP zone. It is important, however, to be aware that the wrong choice of borderline solvent might cause problems. Fortunately, HSP makes it very easy to check whether or not your solvent is in the danger zone.

3.10 The HSP of water

The cohesive energy density of water is high, something we know because of the high latent heat of vapourisation of water. It is quite straightforward to estimate that the δD value should be around 15.5 and δP should certainly be high, say 16. In order for the total cohesive energy to reach the high experimental value, δH has to be 42. A quick check of the δH values of methanol and ethanol (22 and 19) or glycerol (27) makes sense because water is so obviously a strong hydrogen bonder. So the official HSP values for water are [15.5, 16.0, 42.3] and that should be the end of the story.

If, however, you assume that water is a solute and measure its HSP by the classic technique of plotting a sphere with good and bad solvents, using 1% solubility of water in the solvents, then a very different set of HSP values emerges: [15.1, 20.4, 16.5]. You can also get a set of values based on complete miscibility, [18.1, 17.1, 16.9].

How can one solvent have two different values? Presumably, in an alien environment the water molecules can turn in on themselves via internal H-bonds so that the environment sees plenty of polarity but less H-bonding.

Charles Hansen's philosophy is to go with what works. He has found countless examples where the use of the [15.1, 20.4, 16.5] or [18.1, 17.1, 16.9] set produces good, insightful, actionable results, and plenty of other examples where the [15.5, 16.0, 42.3] produces optimal results. In general these different cases fit the idea that water in a minority behaves in the low δH manner and that water in the majority behaves in the high δH manner.

3.11 Are HSP meaningful for large particles?

We already know the pragmatic answer to that question: a resounding yes. Starting with Hansen's own work in the paints industry, HSP are proven to be useful for pigments, nanoparticles etc. All the same, it would be nice to have a theoretical justification for this. Happily, some pioneering theoretical work by the Coleman group in Dublin[34] gives us exactly what we need: formulae for how the dimensionality D affects the formulae for HSP of 0D (small molecules), 1D (e.g. polymers and carbon nanotubes) and 2D (e.g. graphene) materials. It turns out that with no loss of rigour the only difference between the HSP Distance formulae are a factor of (1-D/3), where this is 1 for small molecules, 0.66 for polymers/CNT and 0.33 for graphene.

A significant aspect of this work is that the Coleman group's main interest was not theory for theory's sake but rather in the ability to use and formulate CNT, graphene and other 2D materials. Other papers from the group show the practical utility (with limitations) of HSP for these large, complex materials.

3.12 The problem of small solutes

Moving to a much smaller molecular scale, one of my disappointments with HSP has been that predictions of the solubility of small solutes have been disappointing. Similarly, when I was given privileged access to a gold standard solubility dataset I was surprised and disappointed that attempts to measure the HSP of the solutes, even with more sophisticated fitting approaches using full solubility data, sometimes ended in failure. As I couldn't even define what my problem was, there was no way I could go about fixing it.

Fortunately, a paper by the Rothenberg group in Amsterdam[35] exactly describes this problem, and provides a solution for it: "The [problem] is that the original Hansen parameters exclude thermodynamic considerations. This is acceptable for polymers (where the thermodynamics cancel out) but not for small molecule solutes." The paper includes five ways of improving the ability to measure and predict solubilities. One has been discussed already - handling donor/acceptor effects. Their approach is frank about the difficulties faced by anyone trying to create a set of coherent donor/acceptor parameters and contains some fresh ideas. A second idea involves corrections that allow for yes/no solubility at different concentrations to be accommodated within the fitting procedure. A third way is to invoke a subtle distinction between fitting with the square of the Distance or its square root. A fourth way is to correct for the MPt of the solute, as a higher MPt automatically implies a smaller radius of solubility.

---

34 J. Marguerite Hughes, Damian Aherne, Jonathan N. Coleman, Generalizing Solubility Parameter Theory to Apply to One- and Two-Dimensional Solutes and to Incorporate Dipolar Interactions, J. Appl. Poly. Sci., 127, 2013, 4483–4491

35 Manuel J. Louwerse, Ana Maldonado, Simon Rousseau, Chloe Moreau-Masselon, Bernard Roux, and Gadi Rothenberg, *Revisiting Hansen solubility parameters by including thermodynamics*, XXX ChemPhysChem 10.1002/cphc.201700408

At the heart, and the other corrections feed into this, is the notion of an effective radius, r_{eff}. Instead of just having a Distance and a single sphere radius that defines whether a solvent is good or bad, the Distance is compared to r_{eff} which depends on both the intrinsic sphere radius, r, and a "radius" of the solvent, r_s via $1/r_{eff} = 1/(1/r+1/r_s)$. The authors (private communication) stress that the term "radius" is used only as a short-hand and that in reality they are talking about energy terms.

Conveniently, the solvent radius can be parametrised as $r_s = 1600/MVol$, which means that for a typical solvent of MVol=80 to 120, r_s = 20 to 13. This takes us to their analysis of why no correction is needed for polymers. The reason is that the effective radii are calculated at a standard 0.5 mole fraction. Clearly the radius of a polymer at 0.5 mole fraction is very small (because few solvents will dissolve a polymer by that amount) so the r_{eff} is dominated by the polymer's radius and the solvent correction is irrelevant. This 0.5 mole fraction rule obviously gives problems for high MPt solids and for solids tested at, say, 0.1 mole fraction, hence the need for the second and fourth terms discussed above.

The authors are well-aware of the problems of over-fitting so a proper analysis of prediction versus experiment was carried out, in particular looking for the minimum number of parameters required for good fits. Of their five factors, only the r_s factor is required to get most of the benefit in terms of the quality of fitting *for the training set*. For the set of data *outside the training set* then the predictions required both the r_s factor and the correction of MPt to produce high-quality results. The donor/acceptor correction has a modest effect and the quality of the overall fitting with all five parameters is impressive.

At the time of writing, it remains to be seen whether these important insights can be extended to improve HSP in general or whether they will remain as a valuable tool to be used only for small molecule solutes. In a refreshing change from normal academic practice, the authors have provided their Matlab code (and necessary demo files) so that others can try out the approach for themselves.

3.13 The HSP of nail polish

HSP theory is not too hard, the ideas of how to measure HSP values of industrially-interesting materials are rather simple, and the notion that two bad solvents could create a good one makes intellectual sense. Yet there is a big difference between accepting these ideas and acknowledging that they can work for one's own system. So I am very grateful that the Dutch High Throughput Formulation company, VLCI, invited me to speak at an HSP training day and to watch the attendees do some hands-on HSP experiments. It was fascinating to see attitudes change through the day, with a most interesting flourish at the end.

Finding a system that is safe, quick and reliable for a one-day training course is a big challenge but Sam Peel at VLCI had the inspired idea of using an ordinary nail polish as the "polymer" to be measured. In the days before the course he coated a thin layer of Perfect Touch Rock Chick nail polish onto hundreds of glass slides, resulting in samples with a tough, fairly uniform coating ready for testing within a range of solvents. I am grateful to Sam Peel and VLCI for permission to use the results from that training day in this chapter.

During the first part of the course, the attendees placed the coated slides into tubes of the selected solvents and, after a reasonable time, scored them from 1 (totally dissolved) to 6 (totally untouched). One of the key worries for anyone embarking on their first determination of an HSP is about the reliability of the judgements about the solute being "happy" or "unhappy" in each solvent. The data show that such concerns are, at the same time, valid whilst being largely irrelevant. The solvents include 3 blends (for practical reasons) and the attendees' individual scores are given in columns A-N.

Solvent	A	B	C	D	E	F	G	H	I	J	K	L	M	N
n-Butyl Acetate	2	1	1	2	1	4	4	2	1	3	3	1	1	4
GBL	2	3	2	1	2	4	1	2	2	2	2	3	2	2
DMSO	1	2	5	6	5	2	1	6	2	1	1	1	2	1
MEK	2	4	4	3	4	3	2	3	3	2	2	2	1	1
PGl-Methyl Ether	2	5	3	3	4	2	2	3	2	2	2	2	4	2
1-Hexanol	6	6	6	6	6	6	6	5	6	6	6	6	6	6
Benzyl Benzoate	6	6	5	5	5	6	6	6	6	6	6	6	6	6
Benzyl Alcohol	6	6	6	5	6	6	6	6	4	6	6	6	6	6
Propylene Carb	2	4	2	2	5	4	2	4	3	6	2	4	4	4
2-Propanol	6	6	5	6	6	6	6	6	5	5	6	6	6	6
NMF:DMF 19:81	1	1	1	1	1	1	1	1	1	2	1	1	1	1
CyHex:BrN 48:52	3	5	3	3	3	5	5	4	2	4	2	4	4	3
Hex: n-BA 10:90	3	4	4	4	4	5		2	3	4	3	4	4	5

You can immediately see significant variations, a potential cause of worry, yet when the mode value is taken (effectively eliminating outliers which may have been artefacts from having to prepare so many samples) and the numbers entered into the HSPiP software, the result is clear-cut:

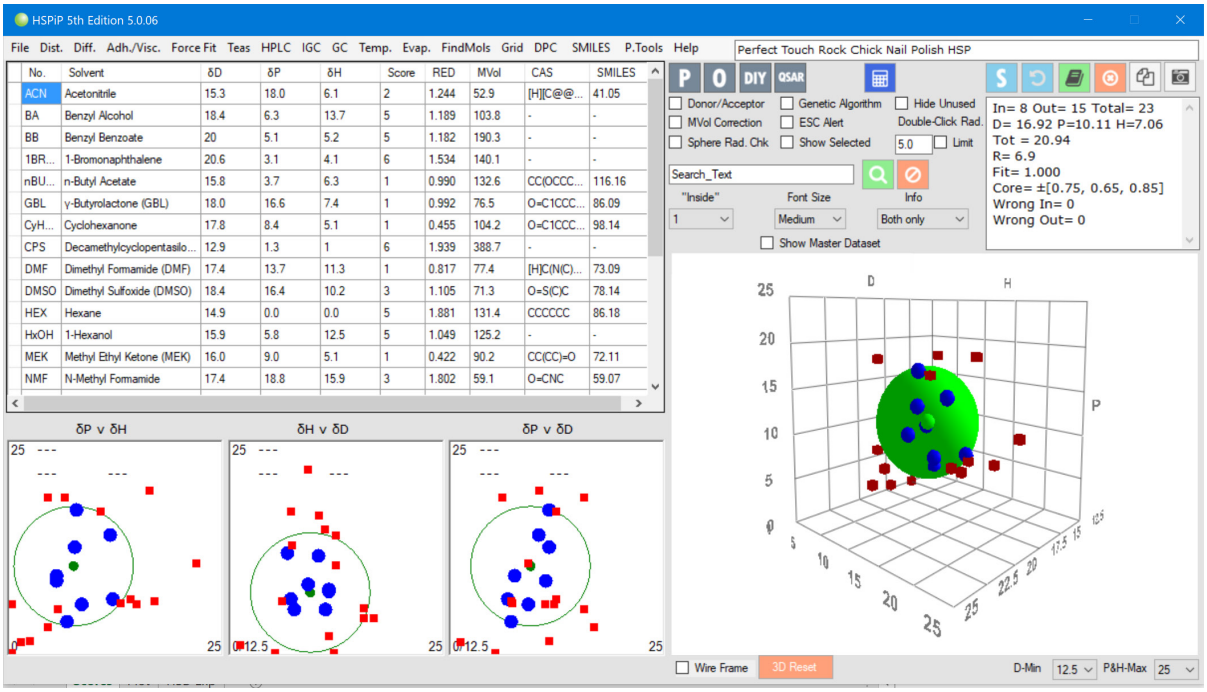


Figure 3-3 Measuring the HSP of nail polish

If those solvents scoring a "1" (really good) are used to obtain the Sphere, the HSP of this nail polish are ~ [17, 10, 7]. If the "2" values are also included then (not shown) the HSP change to [16, 11, 6]. If individual's scores were converted into Spheres, the results were generally not much different. If there had been serious debate about any solvent, and a little more time, then it would have been easy to re-test that one solvent to reach a consensus.

In reality a somewhat larger group of solvents would have been used to reduce the uncertainties. But even in the hands of first-timers working under time pressures, the results are not at all bad. This is typical of HSP determinations.

There is a bonus from this dataset which provided the flourish at the end of the day. There was general agreement that both benzyl alcohol and benzyl benzoate were bad solvents. Yet if you make a 50:50 blend the HSP are inside the Sphere - so it should be an adequate solvent. Each attendee could choose to make a blend of their choice. Those who chose the blend of benzyl alcohol and benzoate found, as predicted, that these two non-solvents became an adequate solvent.

3.14 Working with a new polymer

In 2009 I was asked to write a chapter for a book[36] on the relatively new "green" polymer, poly(lactic acid), PLA. At the time I had not knowingly seen any PLA but it took me less than a day to have a good overview of its solubility properties. From a published list of "yes/no" solubility data it was easy to use HSPiP to fit

36 Steven Abbott, *Chemical Compatibility of PLA: A Practical Framework Using Hansen Solubility Parameters,* Ch.7 of Auras, Lim, Selke, Tsuji (Eds), *Poly(Lactic Acid): Synthesis, Structures, Properties, Processing, and Applications*, Wiley 2010

the data to find the HSP of PLA. Once I knew the HSP I could then address a large number of issues with some confidence. That is when I found that large amounts of work done, with good intentions (because it was a green polymer), were predictable failures. Much unnecessary effort (and precious resource) would have been saved if those working on PLA had taken the trouble to understand its basic solubility properties.

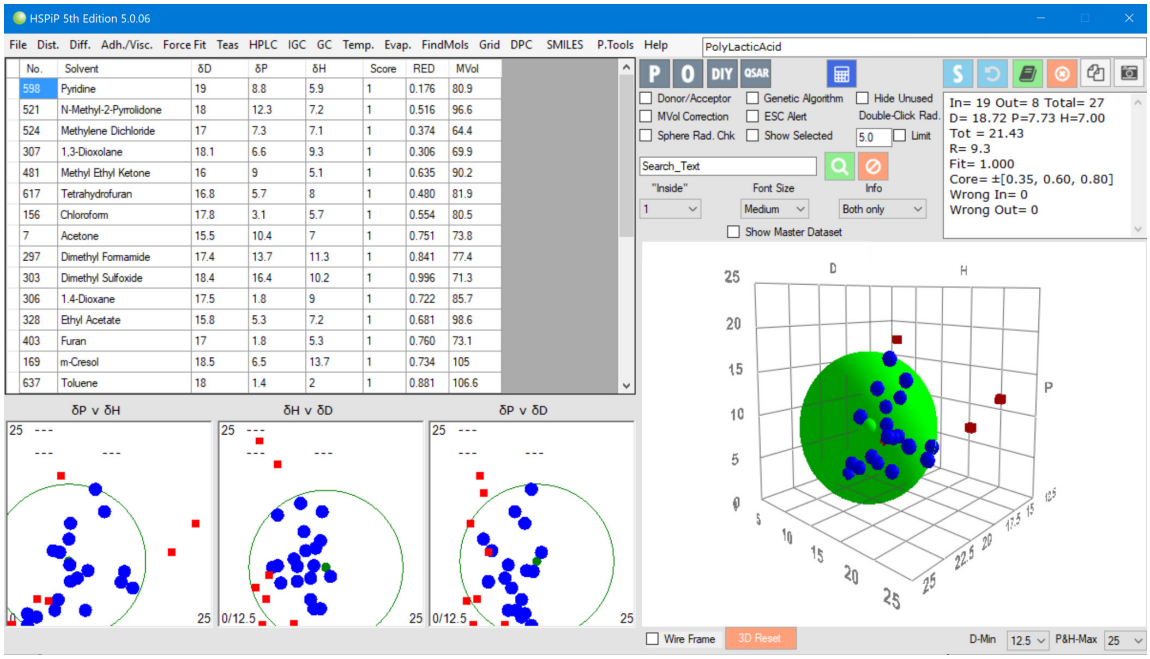


Figure 3-4 The HSP of PLA, from which many predictions could readily be made

A typical paper would say that PLA was a "hydrophobic" polymer (it is insoluble in water) and then propose that it would work with solvents, additives or other polymers to which it is completely unsuited. The first example (of many) was the attempt to use a green plasticiser for this green polymer. The citrate plasticisers are undoubtedly green so these were tried, only to prove a failure over time because the plasticiser gradually migrated to the surface, a sure sign of mutual incompatibility. A few moments with HSP would have shown the futility of using the citrates. With PLA at [18.7, 7.7, 7.0] and tributyl citrate being [16.6, 3.8,10.1], the HSP Distance is a relatively large 6.5. Triacetin has a Distance ~6. One of the benzoates, such as dipropylene glycol dibenzoate [18.0, 6.6, 5.6] has a short distance of 3.5 but is maybe not fully green. With very little effort I could imagine some sort of PLA/PEO combination that would be an excellent match. Sure enough, a patent for such a material claimed excellent compatibility and good plasticisation. Just sitting at my desk I could do in 1 hour what some unfortunate PhDs had failed to do in a few years of work.

Other examples included confident retro-prediction that most tallow-quaternary nanoclays would be hopeless because they lacked any functionality that would make the nanoclay wish to be associated with the PLA. They were justified in the literature because tallows are hydrophobic and PLA is hydrophobic. As expected, when I checked the literature, those nanoclays gave no desirable

properties and the one clay that might have had HSP compatibility contained free -OH groups which would have degraded the PLA during any hot processing step. Using the same world "hydrophobic" to describe a long-chain hydrocarbon (the tallow chains) and a polyester was not a good idea.

I was also able to make the confident prediction that cinnamon buns would not best be stored in PLA packaging because cinnamaldehyde is a close HSP match to PLA. On the other hand, HSP retro-predict that lemon-flavoured goods would be OK because the Distance from limonene is large, a result I found confirmed experimentally.

What I wrote about controlled release for drugs in PLA is discussed in the Diffusion chapter.

It is worth noting a failed prediction I once made about PLA. The issue at hand was the need for a fast-acting, super-safe solvent blend, but one that was near the edge of the solubility sphere. I was able to find such a blend and when tested it was a complete failure. However, when that specific PLA was tested, although its HSP values were (of course) the same as the standard value, the solubility radius was much smaller, 4, compared to "normal" PLA (8) tested under the same conditions. The reason was that the specific PLA was highly crystalline, a fact I had not been told when I made the failed prediction. Using the smaller radius it turned out to be easy to alter the ratio of the original solvent blend, and the system worked as intended.

The sort of methodology described here for addressing broad formulation questions on a new polymer (or new anything) is not something super clever or rare. This is entirely normal for HSP users around the world. It is also not something that is perfectly accurate. Such a simple approach cannot be expected to be right all the time in complex formulations. The strength of the approach is that it can steer you away from the sorts of guaranteed failures that I briefly describe here (and in more detail in the PLA book chapter) so that precious resource can be devoted to formulations less likely to fail.

3.15 HSP via IGC

Although the standard "sphere" technique for measuring HSP values works very well for solid polymers, pigments and dispersions, it is less good for semi-liquid materials such as oligomers, cosmetic excipients or ionic liquids that are rather too miscible with too many solvents so that the sphere is large and ill-defined.

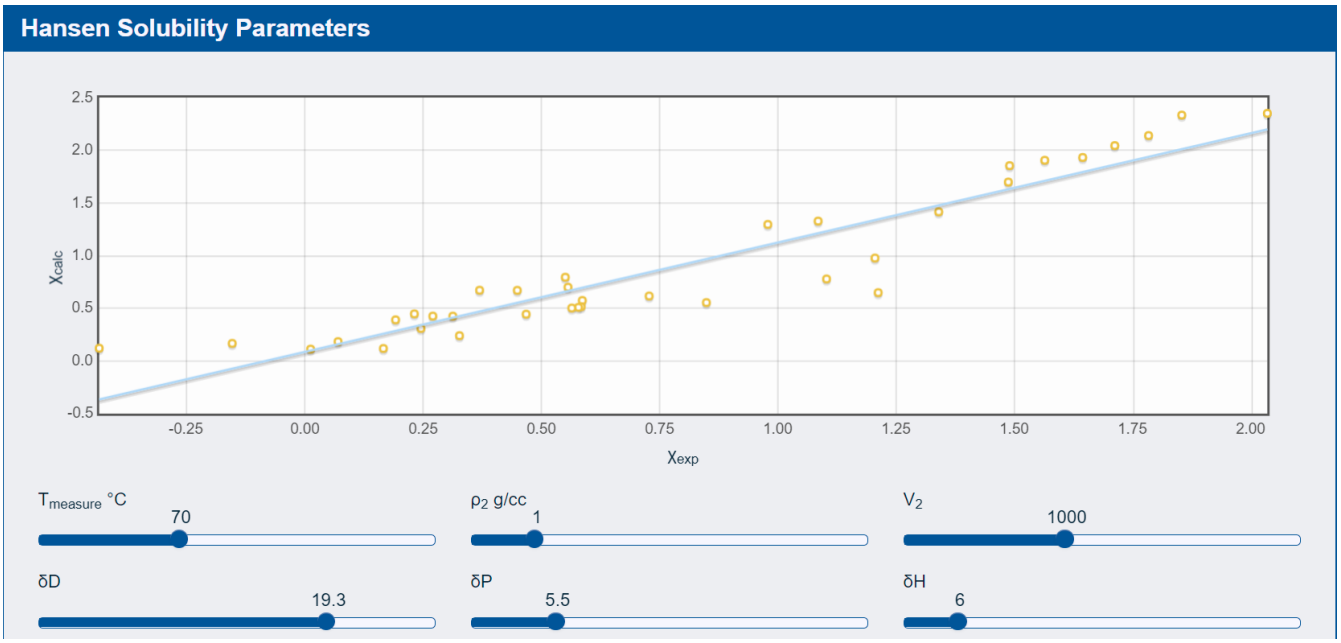
Fortunately, these materials have (provided they are relatively involatile) precisely the right properties for measurement via inverse gas chromatography, IGC. In conventional GC, a standard support material is used to distinguish between the mix of compounds to be analysed. In IGC, the traditional GC column material is coated with a thin layer of the sample material (which is why

it works well with oligomers, excipients and ILs) and then the retention times of a series of known solvents such as acetone or octane are measured. The IGC technique is described in some detail on my Practical Chromatography site, https://www.stevenabbott.co.uk/practical-chromatography/igcbasic.php.

The retention time depends partly on the properties of the probe solvents; low volatility solvents are retained longer than high volatility solvents, so these intrinsic properties need to be taken into account. In addition, the compatibility of the probe solvent with the analyte changes the retention time - more compatible solvents spend more time within the analyte than less compatible ones. The effect is controlled by the χ parameter which, in turn, is related to the HSP distance between the analyte and each probe solvent.

Via a series of steps, described in detail on the IGC app page, it is possible to obtain the best match between the measured χ parameters and the values calculated from the HSP distance.

Although this technique has been known for some time, many of us found that the calculated values were meaningless. Clearly some assumption in the chain of logic was at fault. Work by Dr Eric Brendlé at Adscientis finally proved that the erroneous assumption was that the support material was inert. In fact the different probes interacted with the support materials in different ways, confusing the analysis. With the right support material the artefacts were significantly reduced and measured values usually make a lot of sense. No doubt further refinements will further improve the accuracy of the computed results.



App 3-7 https://www.stevenabbott.co.uk/practical-chromatography/hsp.php

In this example, the data on polycaprolactone from Tian and Munk is fitted with HSP of [19.3, 5.5, 6.0].

3.16 More from the real world

The justification for using HSP is that it is easy to understand and to use, and that it works surprisingly well across a broad range of messy, real world problems. The Hansen Solubility website has a set of such examples (https://www.hansen-solubility.com/HSP-examples/examples-intro.php) to illustrate and to inspire. Some of the topics discussed on the site are:

- **Carbon Black**: How to measure the HSP of carbon blacks, allowing a clear distinction between "hydrophilic" and "hydrophobic" versions.
- **CNT**: How to dissolve/disperse carbon nanotubes or graphene
- **TiO_2**: How to measure/use the HSP of nanoparticles such as TiO_2.
- **Organogelators**: Finding the right balance of solubility for organogelators: neither too low (they crash out of solution) nor too high (they just become solutes).
- **OPV**: Finding the right solvent blends for organophotovoltaic formulations, especially ones that encourage phase separation of the fullerene and polymeric components during the coating/drying process
- **Green Polymers**: How to characterise a new, green polymer to allow efficient selection of plasticisers, nano-clays, drug delivery systems etc. [Discussed above]
- **Skin**: Using HSP for delivery of active ingredients through skin
- **DNA**: Finding the best co-solvent for a DNA hybridisation process
- **Flavour scalping**: Identifying the probability that any given flavour (or mix of flavours) will be lost ("scalped") through food/drink packaging. [Discussed in the Diffusion chapter]
- **QC**: How to characterise different batches of the "same" polymer blend to distinguish between good and bad batches.
- **Cleaning**: Choosing the best solvent (blend) for a cleaning operation
- **Glove safety**: Finding the best glove for resistance to a given solvent or chemical
- **Double sphere**: Identifying if a polymer is best characterised as two separate polymers via a "double sphere"
- **Ceramics**: Optimising a complex ceramic system.

3.17 The future for HSP

One of the many strengths of HSP is that they are simple and can, in principle, all be done with a bunch of spreadsheets, especially because the core Hansen dataset of ~1000 values is public domain. The evidence suggests, however, that the HSPiP software (I acknowledge that as one of the authors I have a potential conflict of interest in saying this) has made a big difference to the usability and applicability of HSP. By bringing together all the standard techniques and adding prediction tools and many sample datasets, users rapidly gain confidence in using HSP and after 9+ years of development most day-to-day problems can be handled naturally within the package. Indeed, the lively HSPiP community (never

shy to send in a bug report or ask for additional functionality) has played an active role in ensuring that the package does what most users want it to do.

Having said that, the limitations of HSP are always made clear. The current absence of a robust donor/acceptor capability is an intellectual and practical concern. The inability to cope with non-mean-field problems is a limitation and HSP have never been fully comfortable in water as an outlier solvent. HSP have no awareness of 3D structures and, as discussed earlier, there are plenty of examples for small molecules where HSP + Ideal Solubility *should* work, but doesn't. HSPiP comes with a large eBook discussing these issues, and users are encouraged to filter HSPiP's predictions through their own ideas of what makes scientific sense. Computer tools exist to supplement our capabilities, not replace them.

The next chapter discusses COSMO-RS. It is indisputable that COSMO-RS is superior to HSP whenever accurate solubility calculations on pure molecules are required. If a low-cost, easy-use version of COSMO-RS were to appear, much of the rationale for using HSP on pure molecules would be removed. But it seems a long path before a COSMO-RS approach could deal routinely with many of the examples in the previous sections.

So HSP, for all its many faults, looks as though it has a robust future. If the latest refinements from Hiroshi Yamamoto around donor/acceptor prove to be sound, or if the "effective radius" approach can be incorporated into HSPiP, these improvements would allow HSP to get closer in quality to COSMO-RS, and that future will be extended somewhat further.

4 COSMO-RS

The theoretical structure behind HSP is full of approximations; the main justification for using it is that it works in many situations, especially with messy real-world formulations where there is no hope of any precise theoretical description. With COSMO-RS we have a solubility theory that is based on a powerful idea with far fewer approximations. For solubility questions that feature relatively well-defined small-molecule solutes it is provably the best tool available as it wins many of the "solubility challenges" designed to test the capabilities of different approaches to solubility. Although that is sufficient to recommend it, COSMO-RS comes with another key advantage - the theory can be readily visualised so the reasons for one solvent being better than another can be understood far better than systems which just provide numbers.

The COnductor Screening MOdel-Realistic Solvation theory starts with a quantum mechanical calculation. This would normally be enough to stop any formulator from going any further. First, quantum calculations are complex and time-consuming. Second, they are in a vacuum which means they are meaningless in terms of real solubility. COSMO-RS solves both problems very elegantly:

- The quantum calculations only have to be done once for any molecule. By now, these have been done for 10's of thousands of molecules and, in particular, for all the usual solvents, so this is no longer an issue. Even for a new molecule, with modern computers the DFT calculations generally take only minutes.
- These calculations are done as if the molecule is surrounded by a virtual conductor ("dielectric continuum"). This means that the molecules have realistic surface polarization charges all around them, not the false charges imposed by a vacuum environment.

These charges can be shown graphically in 3D and already give a strong impression of which solvents may, or may not, interact with which solutes.

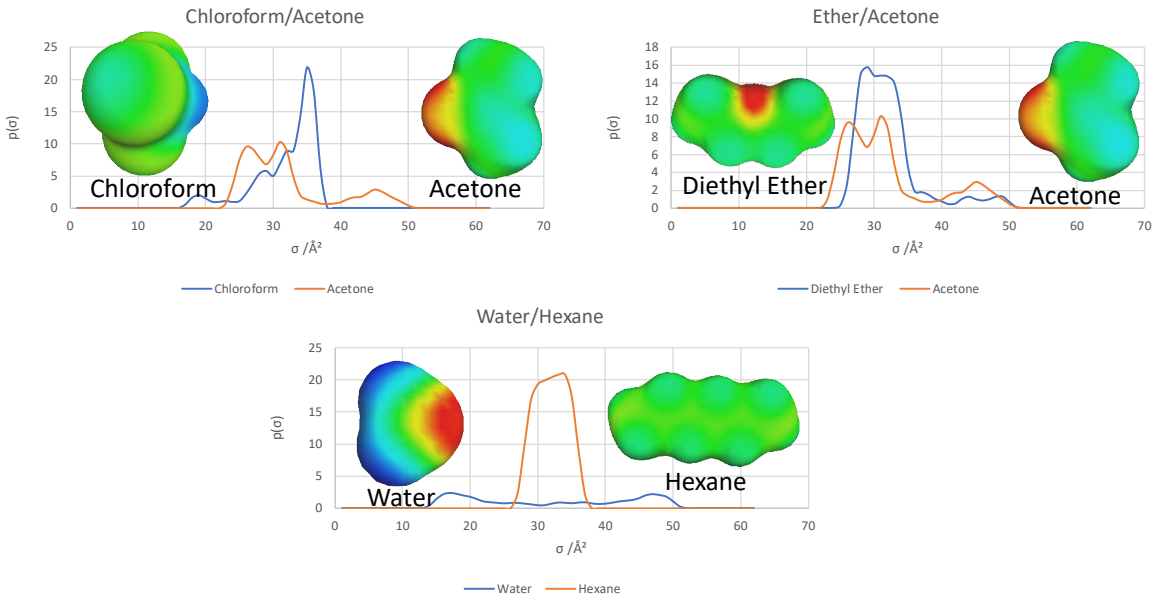


Figure 4-1 The sigma surfaces and σ-profile of three sets of solvents

The positive charge (blue) on the hydrogen of chloroform is rather obviously well set up to match the negative charge on the C=O of acetone, and we know that mixing the two solvents generates heat. Diethyl ether and acetone are clearly rather similar so they do not interact strongly with each other, and we know that their solutions are near-ideal, i.e. their mutual activity coefficients are close to 1. Water and hexane are obviously completely different.

Although the 3D charges, called the sigma surface, provide useful visual clues, they aren't used directly for the solubility calculations. Instead the numbers calculated for the σ-profile graphs are used. The σ-profile is a histogram of the charge intensities around each molecule. For chloroform there is a large peak in the positive domain representing (confusingly, but logically, the negative charge around the molecule) while the hydrogen is the negative peak. For acetone, the interesting peak is the positive one representing the negative carbonyl.

To simplify things greatly, COSMO-RS calculates the interaction energies between molecules by summing the positive-to-negative attractions and positive-to-positive repulsions from putting the surfaces in contact with each other. And because this rather simple summation can be done on any mix of molecules, the process identifies all the interactions in a multi-solvent mix, weighting them depending on the proportions of solvents and solute. So not only can COSMO-RS naturally calculate the thermodynamics of solvent-solute interactions, it can, just like HSP, naturally handle solvent blends. This is of great practical importance.

Some other issues such as van der Waals attractions and entropic terms are dealt with by standard methods.

It is worth emphasising that COSMO-RS is an excellent combination of deep thermodynamics and practicality. Yes, it requires a one-off quantum calculation which is a bit of a nuisance, but then the results of that (slow) calculation can be used for near-instant calculations of the thermodynamics of the solute/solvent(s) interaction. It is far more powerful than HSP because it does everything from first principles (no need to measure the HSP of small-molecule solutes) and provides all relevant information such as activity coefficients or vapour-liquid equilibria.

The theory behind COSMO is precise thermodynamics, but cannot by itself yield sufficiently precise numbers. The RS (Realistic Solvents) part comes via parametrisation of the local polarizability and element-specific interactions (such as dispersion). This required a considerable amount of fitting work against an increasingly large set of high-quality experimental data that has been gathered over the years and the 15 parameters used are now of consistent high quality.

The sigma surfaces shown in the diagrams are straightforward. But what about more complex molecules? This is, indeed, an unwelcome complication. The sigma surfaces of all relevant conformations have to be calculated and used in the solubility calculation process. This in turn means that the conformations themselves have to be deduced, which is not a trivial issue. The relevant proportions of each conformer are calculated via a Boltzmann distribution which, in turn, depends on the interactions which depend on the distribution, so it takes a few cycles to reach a stable solution. In practice, therefore, the issue of multiple conformations is a modest inconvenience rather than a large problem.

This summary of COSMO-RS does not do proper justice to the depth and elegance of the approach. While I appreciate both, they are not the concern of this book. For formulators concerned with pure, smaller-molecule solutes the conclusion is rather simple - use COSMO-RS. Even if its depths and subtleties are not understood, it still works remarkably well. And those who dig deeper into it can tackle more challenging issues. Even so, COSMO-RS has its limitations.

The first is a practical one. It is a big, complex, expensive package which delivers most benefit when it has an expert user with access to a powerful multi-core machine to do the DFT calculations and to handle the generation and calculation of all relevant conformers of any significant solute molecule, plus the ability to mine the wealth of data that is generated from each calculation, and push the boundaries via the various additional capabilities enabled by the COSMO-RS approach.

The second "limitation" (not the fault of COSMO-RS!) is that there is no way to calculate the "ideal solubility" portion of a crystalline solid; so to get absolute solubilities it is necessary either to use the package's estimator, or to provide MPt and ΔH_F data or to give the absolute solubility in one solvent so that all the others could be calculated using that reference. Because the thermodynamics of COSMO-RS includes temperature effects, once the ideal solubility portion

has been defined at one temperature the system is especially useful for identifying the best solvents for purification of (say) pharmaceutical molecules using crystallisation. It can calculate the differential solubilities in each potential solvent over viable crystallisation temperature ranges to pin down the optimum trade-off between absolute and relative solubilities. A solvent with a high high-temperature solubility is not desirable if the low-temperature solubility is also high. A solvent with a huge differential solubility is not desirable if the high-temperature solubility is too low. Finding the right solvent in solubility terms as well as in terms of health and safety, cost etc. is remarkably hard without good solubility predictions.

Another limitation arises because COSMO-RS relies on precise molecular knowledge of the solute so it is hard to apply it to many real-world situations. While this may be an insuperable problem for, say, a nanoparticle or paint pigment, we will explore some recent advances that are helping to overcome some of the issues. First, it is important to see what happens when COSMO-RS is used to tackle the issues of water as a solvent.

4.1 COSMO-RS and Water

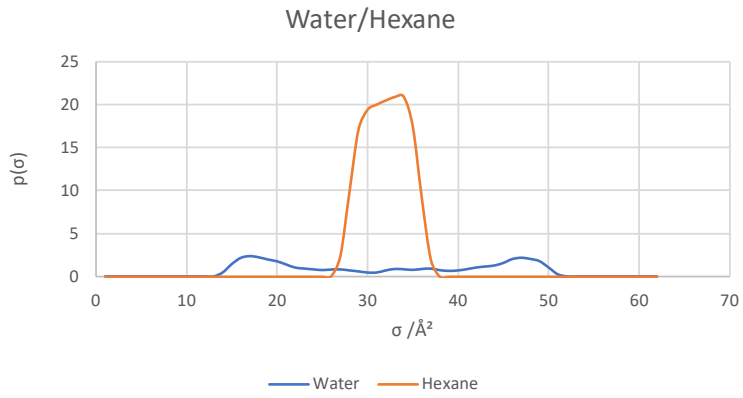
In later chapters we discuss at length the issues around water as a solvent, or, rather, how the perceived mysteries of water as a solvent have created a vast and unhelpful literature and a way of talking about water that frames the wrong debate and adds to the confusion.

It is interesting to note that from the COSMO-RS point of view, water holds (almost) no mysteries and does not require any of the water structures, icebergs and "hydrophobic hydration" effects that (in my view) so confuse what is going on. The reason that COSMO-RS does not require water structure etc. is that it has no mechanism for calculating such effects. Yet it is consistently able to explain, with no special parametrisation, most of the standard issues such as hydrocarbon solubility in water and water solubility in hydrocarbons, in other words it can explain "hydrophobic effects" without invoking anything special.

The "almost" in the previous paragraph exposes one acknowledged weakness of the approach. There is a special "surface scaling" parameter applied uniquely to water's sigma surface (making its surface effectively smaller) without which all activity coefficients are too large. And in the case of the extremely delicate balances (discussed later at length) behind Lower Critical Solubility Temperatures, it is necessary to make this parameter adjustable. In most cases, therefore, the "special" nature of water is accommodated by a single ad hoc parameter which has been built in to the software.

How can COSMO-RS produce objectively good results when all it has is statistical pair-wise interactions between molecules? The answer is "via statistical pair-wise interactions between molecules". That is all that is required.

And to emphasise the "visual" nature of COSMO-RS we can see with our own eyes why this is the case.



When plotted against the σ-profile of hexane, water has very little σ-surface because it is a small molecule. And the surface that it has is concentrated in the (charged) wings either side of the neutral centre. Although, like all molecules, water has a neutral portion, it is very small. So when you try the pair-wise statistical overlap of the water with the hexane, and compare it to the overlap of hexane-hexane and water-water, the statistical attraction is tiny.

Here is where things get interesting. In enthalpic terms, hexane doesn't much care whether it is or is not in water. Surprising but true. The problem is that only a statistically small number of the possible water-hexane pair-wise interactions is enthalpically neutral. We get enthalpic neutrality at the cost of a statistically unlikely set of pair-wise interactions. Another phrase for "statistically unlikely" is "high entropy". It is the *entropic* penalty of putting the hexane into water that causes the problem. Note that there is a big difference between this notion of the entropic effect compared to the "water structure" or "cavity formation" or "iceberg" type explanations.

And it remains interesting, because the water solubility in hexane has a different explanation. Here the enthalpic penalty of removing water from water is huge. But those water molecules forced to be near the hexane don't care about their orientation compared to their partial restriction within water. So the huge enthalpic penalty is half-compensated by the large entropy gain.

I find this de-mythologised version of water's interaction with hydrophobic structures most helpful. This is because in the early KB chapter and also in the later water chapters we will find that most interesting effects (salts, urea, etc.) are best described by specific local interactions and *not* by the generalised water-structure interactions that are most commonly proposed. COSMO-RS is an existence proof that nothing is required other than statistical pair-wise interactions, totally ignorant of extra structures.

To finish this section on water, I want to compare and contrast two very similar molecules to emphasise that relatively small effects can have large effects in terms of solubility.

The two molecules are: t-butanol, t-BuOH or TBA and trimethylamine N-oxide , TMAO or TMNO. The first is the only butanol which is fully miscible with water in all proportions and temperatures. The second is a "hydrotrope" or "osmolyte" used especially by deep sea fish to stop their proteins from collapsing under high pressure[37]. The reason for comparing them is that they both have the large hydrophobic trimethyl group, they are very much the same size and they can both hydrogen bond - and yet their aqueous solubility behaviours are totally different. In a later chapter I mention a molecular dynamics comparison of the two. Here I just show a comparison derived from COSMO-RS calculations.

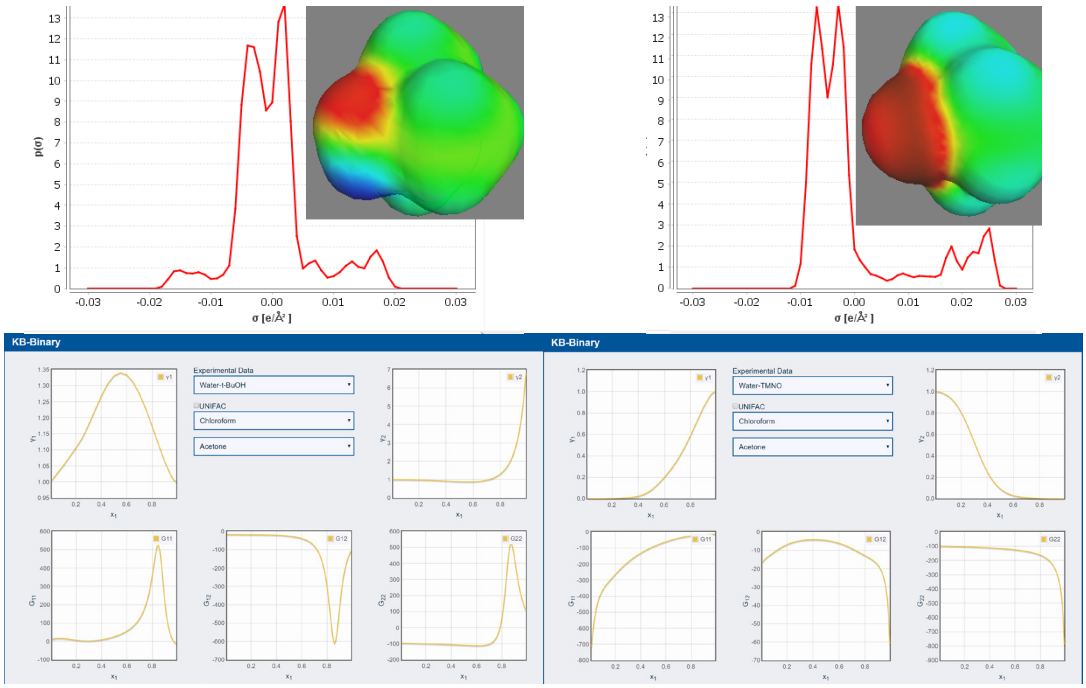


Figure 4-2 Compare and contrast t-BuOH and TMAO with COSMO-RS and KB

The t-BuOH on the left shows a potential advantage in that it has the ability to interact with water as an H-bond donor and acceptor. The two "wings" in the σ-profile match those of water. TMAO on the right shows a high charge density, so the wing to the right is somewhat larger and bigger. The central elements are very similar, a typical large hydrophobic entity.

The point of showing the comparison is that relatively small differences can have large effects. The full KB analysis of water-t-BuOH and water-TMAO can be found in the Binary KB app, https://www.stevenabbott.co.uk/practical-solubility/kb-binary.php. They are in the "experimental data" section though I created the activity coefficient curves via COSMO-RS. The t-BuOH G_{ij} match with experimental data (discussed later) from scattering is impressive. Although

37 TMAO is good at keeping proteins in their folded, compact form, via excluded volume, i.e. it does not interact with proteins. So how can it stabilise deep-sea proteins, surely it would encourage further compaction by crushing? It turns out (there is a paper by Shimizu and Smith) that TMAO happens to be more excluded from the crushed state than from the folded state. Urea, interestingly, destabilises in *both* directions: it unfolds proteins at normal pressures and encourages crushing at high pressures. Why? The KBI say what is interacting with what, the deep reasons for the numbers will be the usual balance of many small effects.

t-BuOH is soluble in water, the G_{ij} values show that it is deeply unhappy. Not shown in the screenshot is the excess numbers which reach 15 for water and 2 for t-BuOH. Statistically they like being away from each other. The TMAO activity coefficients are unrealistic but the general trend is correct; TMAO is not only happy in water but they positively prefer to be together.

I had always thought that mutual solubility or immiscibility were clear yes/no distinctions which would be driven by large and obvious forces. The opposite is the case. Even a package as powerful as COSMO-RS cannot reliably predict aqueous miscibilities and their temperature dependencies. This is because just 1kJ/mol in either direction can make a difference. For example t-BuOH is entirely miscible with water while 2-butoxyethanol is miscible at lower temperatures but is immiscible over 50°C. When (data not shown) I look at the sigma surfaces and σ-profiles, or the output data of the solubility behaviour of the two molecules, there is not a wild difference between them. The very different behaviours depend on the balance of small effects. These two molecules will be discussed further in the Language chapter.

When people talk about some molecule behaving in a "special" manner in water, what this usually means is that a few kJ/mol either way is enough to tip the balance. There is nothing special; it's just that's how the balance has worked out.

That sounds either glib or depressing. I find it liberating. I also find that KB helps put these effects into context. Here, again from the binary mixes app, are the G_{12} values for methanol, ethanol, propanol and t-BuOH. The other butanols are *not* shown because they happen (by a kJ/mol or two), to be insoluble.

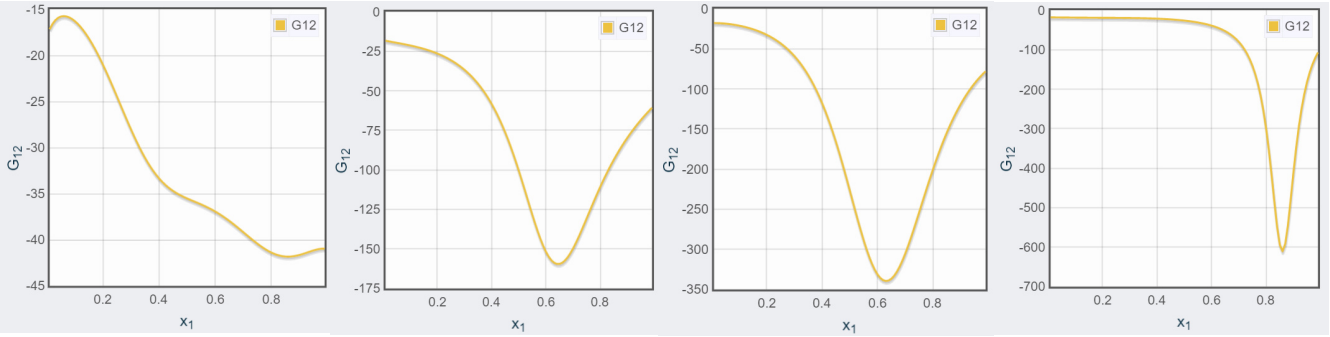


Figure 4-3 From left to right the G_{12} values for water with methanol, ethanol, propanol and t-BuOH

It is instantly clear that the mutual unhappiness (negative G_{12} values) increases along the series. Exploring the app shows everything else that is going on. It is remarkable how much molecular knowledge is gathered starting from just a set of rather dull activity coefficient curves. The fact that the large negative peak in t-BuOH is far over to the right is simply because the plot is in terms of mole fraction which is distorted by the large difference in the MVols of the two

components; the peak is, as you would expect, in the 50% volume fraction range, just like the others.

4.2 Beyond simple solutes

It is obviously impractical and unnecessary to do a quantum calculation on a whole polymer (and its many conformations) in order to calculate polymer solubility. Using calculations on sub-sections of a polymer (with some tricks to "cap" the end then to computationally ignore the caps) it is possible[38] to create the σ-profile representing the essence of a polymer and from that the solubility properties can be calculated. The entropic term is somewhat more difficult to handle but methods are available. The predictive power of this approach in terms of solubilities of gases and liquids in the selected polymers is impressive. In particular the classic issue of partition of fragrance molecules between water and polymer can be predicted satisfactorily and this is key for "flavour scalping" between a food/drink and the packaging around it. Although it might seem straightforward to apply the approach to solubility of polymers in solvents (the other side of the spinodal), it has so far proven to be rather difficult.

When we come to ionic liquids (ILs) there is no lack of effort within the COSMO-RS community to allow meaningful calculations of these difficult materials and there is impressive success with small molecule solubilities. Because (as discussed in the "green" chapter) it is unlikely that ILs will find much use in conventional solubility problems, ILs will tend to be used in areas where their unique properties enable new possibilities. Two examples give a flavour of how COSMO-RS can provide (related) insights into tricky solubility issues.

The first is the solubility of large flavonoid molecules which have features of multi-functional phenols and of sugar-like molecules. Many of these molecules are insoluble in conventional solvents because all those phenolic and sugar -OH groups can form strong inter-molecular bonds. To dissolve them requires a solvent that interacts more strongly with those -OH groups than the molecules themselves. Some ILs can do this very well and a virtual screening[39] of over 1800 ILs revealed two rather simple rules of thumb. The first is that the cations do not make a lot of difference. The second is that "harder" anions, things as simple as Cl^- or acetate, are better than the big, complicated anions that often feature in IL studies. There is a third idea that is plausible but un-proven: water (e.g. contaminating some acetate ILs) can so attract the anion that the anion loses its effectiveness with the flavonoid.

---

38 Christoph Loschen and Andreas Klamt, *Prediction of Solubilities and Partition Coefficients in Polymers Using COSMO-RS*, Ind. Eng. Chem. Res. 2014, 53, 11478−11487

39 Zheng Guo, Bena-Marie Lue, Kaj Thomasen, Anne S. Meyer and Xuebing Xu, *Predictions of flavonoid solubility in ionic liquids by COSMO-RS: experimental verification, structural elucidation, and solvation characterization*, Green Chem., 2007, 9, 1362–1373

The second example is cellulose where most conventional solvents are useless and where ILs can be excellent solvents (at least at high-enough temperatures to give low-enough viscosities). Celluloses are even clearer examples of polymers that should be trivially water soluble because they are just chains of sugars, yet the internal H-bonding is so strong that water can merely swell the cellulose, not dissolve it. There are a number of papers on the topic of COSMO-RS, ILs and cellulose but the one I wish to discuss[40] includes a very striking vindication of the point that COSMO-RS not only provides numerical results but also provides visual intuition. How does one model cellulose within COSMO-RS? A plausible approach is to take cellotriose, where the central sugar is representative of cellulose, do the quantum calculations on the different conformers then (as discussed a few paragraphs earlier) use a trick to block out the two sugars at the end when calculating the sigma surface and σ-profile. What is striking about the calculations is that the σ-profiles for conformations with and without internal H-bonds are so significantly different. It is obvious to the human eye that the conformations without H-bonds will interact much more strongly with H-bond acceptors. We have always known that this must be the case, but it's exciting to see it and get an idea of how this might play out in terms of solubility.

Well, that isn't true. COSMO-RS can only show the internal H-bonds which aren't the reason that cellulose is insoluble. The barrier to solubility is the bonds between molecules. Nevertheless, COSMO-RS must be giving us a reasonable insight into the real situation, because the results of the simulation map quite well onto the available experimental data.

And what do we learn? The authors themselves point out the similarities to the flavonoid paper: cations make relatively little difference because they have little chance of interacting with the H-bonds; "harder" anions are more effective at solubilising cellulose, as is found experimentally.

The flavonoid paper uses some of the wealth of internally-generated COSMO-RS data (such as the energies specific to H-bonds) to check that the calculated effects are primarily due to the stronger H-bonds with the harder anions. Sure enough, there is a strong correlation between the calculated H-bonding interaction energy and solubility, and there is no correlation with van der Waals or the so-called "misfit interaction" energies.

Although these general trends are insightful and convincing, there are plenty of problems and issues relating both to COSMO-RS (e.g. how best to represent the ILs within the calculation) and to the ILs (especially the effects of water). The further the systems get away from already complex systems such as the

---

40 Jens Kahlen, Kai Masuch and Kai Leonhard, *Modelling cellulose solubilities in ionic liquids using COSMO-RS*, Green Chem., 2010, 12, 2172–2181

flavonoids or the cellotriose, the more complex will be the calculations needed to deal with messy reality.

Returning to a more general theme, if COSMO-RS has some difficulties with polymers, it requires heroic efforts to be used for systems such as pigments or nanoparticles. Similarly, the world of cosmetic formulations with complex multi-component systems is hard to combine with the elegant purity of COSMO-RS.

4.3 The future of COSMO-RS

I am not aware of any other solubility theory that comes close to the power of COSMO-RS. While many of us might hope that it was rather cheaper and rather more user-friendly, we also acknowledge that it has to be a big, complex package to cope with the wide range of demands placed on it. As just one example, in this book I have no interest in vapour-liquid equilibria, whereas many chemical engineers absolutely require the predictive power of COSMO-RS in this area.

The COSMO-RS user community is lively and demanding and keeps pushing the boundaries of the capabilities of the approach. So the package continues to develop both in the core functionality and in its complementary add-on packages. If, in time, it can also deal with surfactants, hydrotropes and solubilizers its power will be truly awesome.

5 Dispersions

Where is the border between an entity being "soluble" and being "a dispersion"? Remarkably, we know the answer: there isn't a border. By applying the formal criteria of fundamental thermodynamics via the Gibbs Phase rule, it is clear that there is no difference.

[Skip this paragraph if you wish. The argument, from Shimizu and Matubayasi[41] goes like this. If we have two components (C=2) which are an entity (solute or particle) as a single phase (P=1) in a solvent then via Gibbs Phase Rule you have F degrees of freedom given by C-P+2=3. This means that you can change temperature, pressure and composition arbitrarily. This is a definition of a solution so the fact that one of the entities is a particle is irrelevant - it is still a solution. As far as I know, most dispersions allow such arbitrary change of any/all of the three parameters, so their behaviour can be described by solubility theories, especially including KB. Another approach is to measure the D parameter (discussed in Chapter 1) in KB theory $D=x_2(\delta\mu_2/\delta x_2)$ via neutron scattering. If D>0 then again we have a solution which does not care if the solute is a molecule or a particle. This is deep science but is immensely powerful for those who care to explore its implications further.]

And using informal criteria there is also no difference. For example, "dispersions" are large particles. Yet although a large strand of DNA is visible with a microscope, it is accepted as being soluble. Dispersions tend to flocculate. Yet proteins which we know to be soluble can readily flocculate. Dispersed particles tend to sink with time to the bottom of a test tube. Yet proteins (again) can readily be pushed to the bottom of a centrifuge tube by the high g-forces; so the fact that one thing, "a dispersion", can fall to the bottom with 1g whereas another thing, "a solute", requires 100g is not a criterion for distinction. And in any case, nanoparticles with a size small enough to be bounced around by Brownian motion will not settle even if a slightly larger version of the same nanoparticle does settle. The general rule of thumb is that particles of radius r and density relative to the solvent of $\Delta\rho$ will not settle if $r^4g\Delta\rho<kT$ where kT is the Boltzmann constant times temperature, a value that appears regularly in this chapter. An app that calculates standard Stokes settling times along with a "is it too small to settle?" calculation can be found at https://www.stevenabbott.co.uk/practical-solubility/stokes.php.

Although there is no reason why standard solubility theory cannot be applied to dispersions, at the time of writing, the literature on applying KB to dispersions is very slight. I predict that the situation will change significantly over the coming years with KB ideas becoming one routine way to tackle some complex issues. As with proteins, the excluded volume effects will be dominant and yet they are

41 The short summary is courtesy of Dr Shimizu and is backed up by a number of their papers for those who want to dig deeper.

currently little understood. Many strange dispersion effects will seem far less strange once excluded volume becomes a natural part of the language.

Finally, we know that HSP work very well with many pigments and nanoparticles and the Coleman paper cited earlier fits large particles into standard lattice theory and HSP theory.

So we have multiple reasons to think of dispersions as "solutions".

Still, it is convenient to use some standard "dispersion" ideas on larger particles. So this chapter in a "solubility" book is unashamedly about these "dispersion" ideas. As we shall find, unsurprisingly, some of these dispersion ideas are directly related to solubility ideas.

5.1 DLVO

It is not too hard to imagine that a theory of the stability of dispersions should take into account three effects:

1. van der Waals attractions which mean that all particles want to clump together;
2. repulsions between particles if they are charged;
3. steric repulsion from polymer chains sticking out from the particles.

Since the 1940s one theory has done this[42] and still seems to have no practical rival. It is called DLVO, named after Derjaguin and Landau who developed the theory in 1941 and Verwey and Overbeek who were unaware of the Russian literature and came up with the same ideas in 1948.

There is general agreement that DLVO theory suffers from two major problems. The first is that it is a great over-simplification and the second is that it is way too complicated with too many variables that are unknown in practice. The famous quote from Christenson[43] puts this double problem rather nicely: "... DLVO-theory is completely inadequate (to put it gently) in almost every system so far investigated". One of the reasons for Christenson's critique is discussed in the Hofmeister section of the Aqueous Solubility chapter. The title alone talks of solvation, hydration and capillary effects as being "non-DLVO forces", so clearly DLVO has many flaws.

There is also general agreement that despite these problems, formulators who grasp the DLVO principles are more likely to succeed than those who don't. There are, of course, "Extended DLVO" theories that attempt to fix some of the

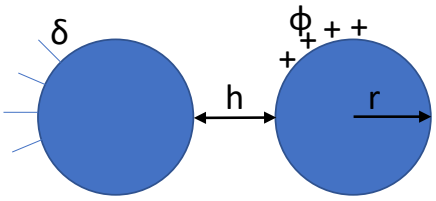
---

42 Original DLVO did not include the steric effect, but the overall narrative is much simpler if I assume that it did.

43 Hugo K. Christenson, *Non-DLVO Forces Between Surfaces - Solvation, Hydration and Capillary Effects*, J. Dispersion Science and Technology, 9, 1988, 171-206

many issues. If these extended theories provided lots of benefits then we would all be using them, and we're not.

The approach adopted here is to present the formulae for each of the three effects, then to identify the key inputs required for a calculation, then to provide users with the app that handles the calculations. This is pragmatic DLVO theory - the maximum benefit for the minimum work. Those who want a deeper understanding of any of the components can readily find explanations on the Internet.



The key parameters in each of the three equations are the radius of the particle, r, and the distance between particles, h. The diagram also shows δ, the length of the chains sticking out from a particle and φ, the charge on the particle.

The first term in DLVO is the van der Waals term though I have called it V_H because the strengths of the van der Waals attractions between two particles are calculated using Hamaker's method and the Hamaker constant A_{12}. The calculations are all done in units of kT, Boltzmann constant times (absolute) T:

Equ. 5-1

$$V_H = -\frac{A_{12} r}{12 h k T}$$

Larger particles have a stronger self-attraction and as h becomes very small, V_H becomes enormous - i.e. if the particles ever get close-enough to touch, nothing can separate them. This reiterates the point mentioned in the HSP chapter that van der Waals forces seem very dull, yet the fact that everything is attracted to everything else via van der Waals means that they are the single most important force in solubility. What are the values of the Hamaker constants for a given system? They can be measured with great difficulty but usually it is not worth the bother; a value of $1E^{-20}$J (or 10 zJ for zeptoJoules) is good enough for most of us. A table of Hamaker values is given in the app.

The second term is the charge-dependent term that I have called V_D because it is based on the Debye-Hückel calculation of charges within ionic solutions. The formula is complex and contains the confusingly-designated k^{-1} parameter which is not to be confused with the k in kT. It also includes Avogadro's number N_A:

Equ. 5-2

$$V_D = -\frac{2\pi e_0 \varepsilon r \varphi^2 \ln(1 + e^{-\frac{h}{k^{-1}}})}{kT}$$

$$k^{-1} = \sqrt{\frac{e_0 \varepsilon k T}{2 N_A e^2 I}}$$

Equ. 5-3

Fortunately the app deals with all this complexity. Just for reference, here is what the parameters mean. The k^{-1} is the Debye screening length (how quickly the effect of the charge disappears over distance) and depends on e_0 which is the permittivity of free space, e is the charge on an electron, ε is the dielectric constant of the medium and I is the ionic strength which (as per the app) depends on the concentration of salts in the solution and their charges Z_1 and Z_2 which would each be 1 and 1 for NaCl and 2 and 1 for $MgCl_2$ and so on.

The dependence on h is relatively gentle and long-range so if there is a significant charge φ, a low ionic strength I and a high dielectric constant ε there is a strong repulsion keeping things apart. So large salt concentrations are bad, any solvent other than water (ε=80) is bad and, of course, you need a lot of charge on the particle. What, then, is φ? Embarrassingly, the answer is that no one really knows, but if we use the measured zeta potential, ζ, that seems to be good enough. When we explore ζ you will see why there is so much vagueness.

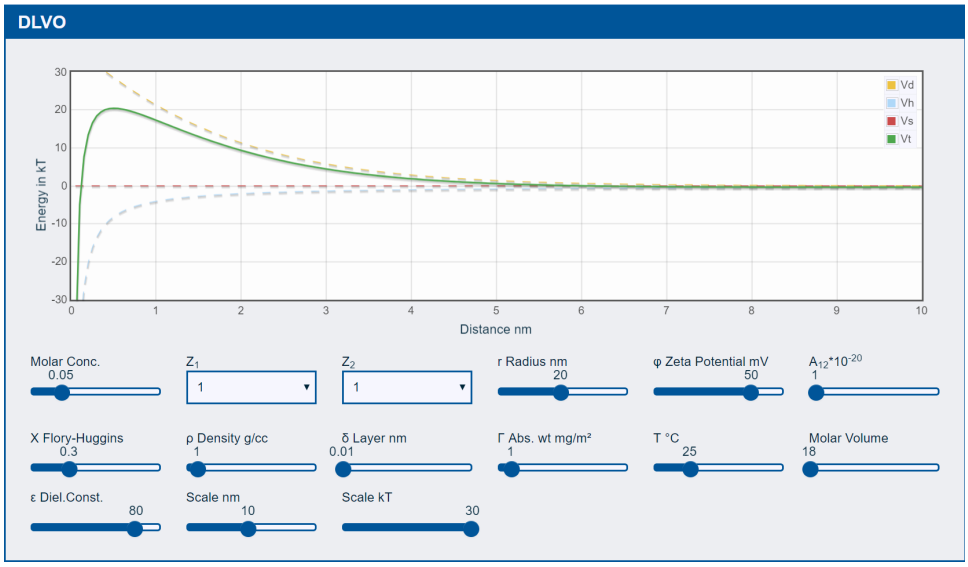
Finally we have the steric term, V_S. Here we use the value of δ, the length of the molecule (we'll call it a polymer for simplicity though it could be a long surfactant chain) sticking out from the particle plus Γ which is the coverage of the particle by the steric barrier material, and a density ρ and MVol that are of no great significance. The fact that the χ parameter (as discussed in the HSP chapter) appears is evidence that "solubility" and "dispersions" are, indeed, related:

$$V_S = \frac{30 N_A \pi r \Gamma^2}{\rho^2 MVol} (0.5 - \chi) \left(1 - \frac{h}{2\delta}\right)^2$$

Equ. 5-4

The equation is only valid when h>2δ for the obvious reason that there is no steric force until the steric chains are touching. V_S suddenly becomes very large and positive (repulsive) which is why steric stabilisation is so popular. However, there is a catch - that term with χ inside. As we know from previous chapters, χ captures how (dis)similar solvent and polymer are. If they are identical (in terms of interactions with themselves and each other) the χ=0 so the (0.5-χ) term has a maximum value of 0.5. As solvent and polymer become more dissimilar, χ increases. If you try doing this in the app, not a lot happens because the V_S term is so large that decreasing it to, say, 0.25 when χ=0.25 is not significant. Indeed, a formulator can unknowingly end up with a solvent that takes χ close to 0.5 and still see no problem. The problem arises when the solvent changes to slightly above 0.5. Now the (0.5-χ) term is negative and V_S becomes negative (attractive), causing the whole dispersion to coagulate. Many of us have suffered from this large effect from a small changes without knowing why. Hopefully, knowing the root cause will help formulators to be in a safe, low-χ zone.

Now we can bring all the terms together, giving a total potential, $V_T = V_H + V_D + V_S$. The aim is to keep V_T as positive (repulsive) as possible for as short a distance as possible so that particles that happen to be approaching each other will bounce off rather than stick together. We can now discuss why V_T is expressed in terms of kT. The average particle will be moving with that Boltzmann energy of kT with fewer (exponentially) at higher kT values. The rule of thumb is for the barrier to be 20kT which means a probability of e^{-20} of overcoming the barrier, i.e. 1 in a trillion. If the balance of forces gives a minimum of a few kT at some significant distance, that indicates that the particles might pseudo-clump, which may be worrying but which can usually be overcome by re-mixing/shaking.



App 5-1 https://www.stevenabbott.co.uk/practical-solubility/dlvo.php

The first example shows a typical charge stabilised system with a zeta potential of 50mV in water and a suitably low molar concentration of ions. The particles reach a barrier of 20kT at around 0.5nm, perhaps dangerously close if there are imperfections somewhere in the system.

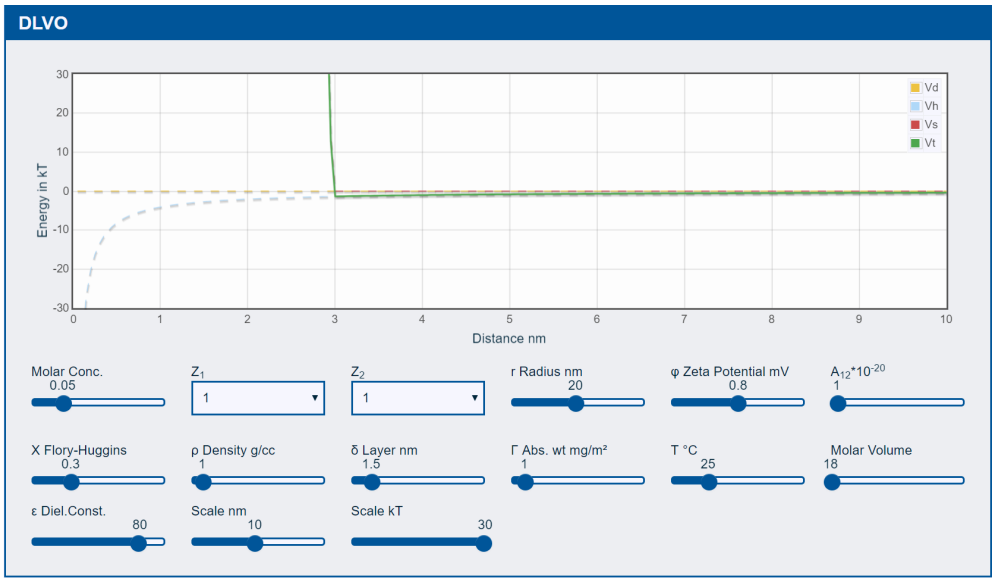


Figure 5-1 Pure steric stabilisation for the same system

The second example shows an uncharged particle relying on a 1.5nm protective shell, leading to stable distance of 3nm.

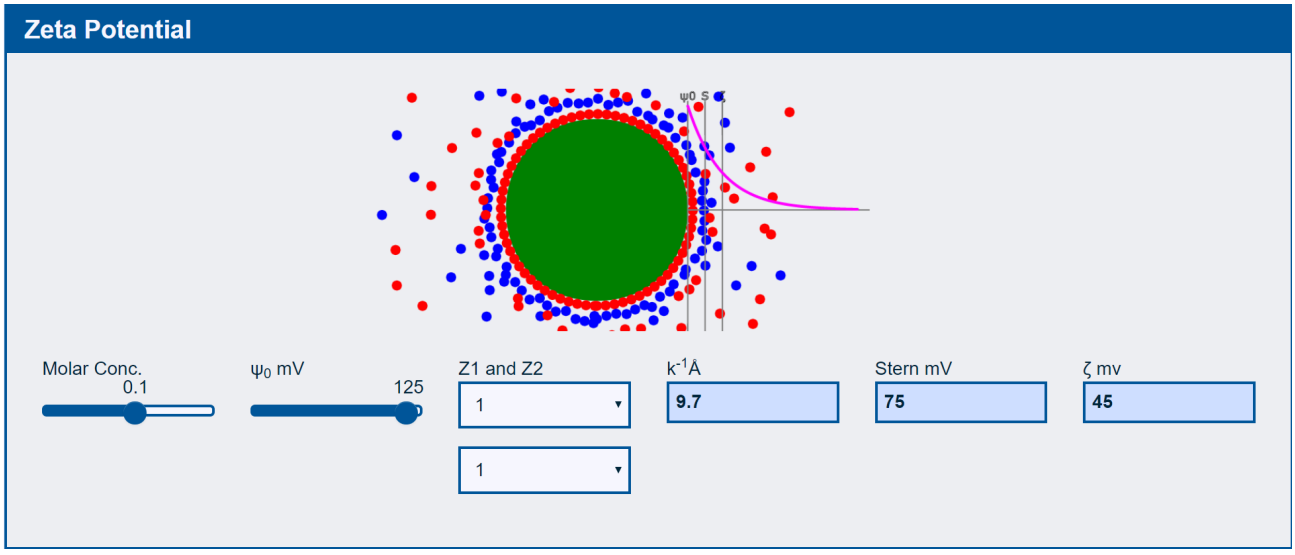
The key advice for handling the app is to build up confidence by leaving all parameters alone except for r, φ and δ. Playing with these gives a good feel for the key effects, the shape of the V_T curve and the relative chances of clumping (barrier <20kT), pseudo-clumping (a minimum of a few kT) or general stability at a safe distance (high φ) or at the 2δ distance where strong steric stabilisation kicks in. Those who rely on charges should then reduce the dielectric constant from water's value of 80 to that of common alcohols around 30 to see that water is probably the only environment where charge-based stability is viable.

For those choosing to rely on steric stabilisation, the question is why it works at all. Imagine a sphere of radius r which self-associated via van der Waals attraction. Now put a really good polymer layer around it so that its radius is r+δ. Why doesn't this clump because of van der Waals attractions between the polymers? That is a very good question and it turns out that people are still debating the issue. My personal favourite is that it is an "excluded volume" effect in terms of KB theory. This means that the KBI of two large molecules get off to a bad start because they can't begin to have mutual interactions till the radius in the KBI calculation is larger than those large molecules - compared (we always need a comparison) to the small solvent molecules which start their KBI with respect to the large molecule with less of a disadvantage. It turns out that "excluded volume" effects and "entropic" effects and, even, "osmotic pressure" effects are basically the same thing, seen from different viewpoints. The classic osmotic pressure explanation for steric stabilisation states that the local polymer concentration is larger if two particles get close and solvent needs to flood in to equalise concentrations. Unfortunately, none of these equivalent theories explain why the χ parameter is critical. The simplest explanation I can find is that the excluded volume/osmotic effects are fine if the polymer-solvent interactions aren't too different from polymer-polymer interactions, but as soon as the polymer is no longer happy in the solvent (which happens by definition at the θ point when χ=0.5) then we are back to the r+δ ball which just sticks together via van der Waals.

We do not have to care too much about the detailed explanations as they are concerned with idealised systems and we have messy real-world systems. We *do* need to care that the 2δ protection distance and the sudden clumping when χ>0.5 are well-validated effects in model systems and are good-enough principles for formulation work.

5.2 Zeta potential ζ

The charge term φ in DLVO is rather complicated. So the convention is to use the measured ζ potential. A typical lab particle-sizer which provides r is also able to apply an electrical field across the measurement cell and from the way that the particles move in that field, ζ can be calculated. Given that the measurement is simple and pragmatic, it is good enough for most of us. Unfortunately, although one might think that "the charge on a particle" is straightforward and self-evident, experience shows that the value of the ζ potential even of a relatively simple particle such as silica can vary wildly not only with obvious things like pH but with subtle things like low levels of ionic impurities. It is, therefore, important to see why "the charge on a particle" is not a simple concept. The app shows this nicely.



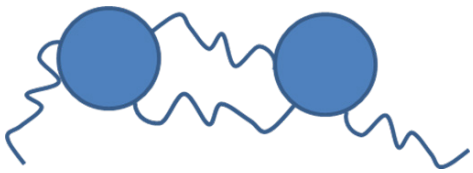
App 5-2 https://www.stevenabbott.co.uk/practical-solubility/zeta.php

The input value ψ_0 represents the "real" charge on the particle, though this is a meaningless concept. This real charge is the ring of red dots around the particle and it attracts, naturally enough, a ring of negative charges in blue. And we can immediately see why "the charge" is such a tricky concept. If that ring of negative charges were, in general, perfect then the charge of all particles would be exactly zero. That is clearly not what happens, but undeniably the measured voltage will be significantly less that the "real" charge. And, of course, that imperfect ring of negative charges will attract an even more imperfect ring of positive charges with an even more uncertain impact on the final charge. The usual hand-waving way out of this is to say that the first pair of charges forms the Stern Layer and that at a somewhat larger radius you have the swarm of charges that, on average, move along with the particle, as opposed to those ions that are sitting around in the bulk solution that do not move if the particle moves. The point where the moving cloud ends and the bulk solution begins is where the ζ potential is "measured".

Now you can see how relatively small effects can have amplified effects via their impact on these vague clouds of particles. Some large multi-valent ions that happen to be near the particle, or some contaminants at the true surface of the particle can each be imagined to have complex effects on the ionic cloud and, therefore, on the measured potential.

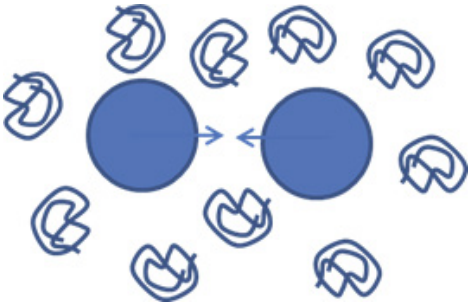
This rather jaundiced way of looking at ζ potential is deliberate because it leads to the key rule about ζ potentials: "Never measure *a single* ζ potential, measure the potential in a variety of conditions relevant to your formulation". In the past this would have been unwelcome news because measurement was so hard. With modern equipment there is no excuse for not following this rule. Suppose you deliberately change the pH across the range of any likely formulations. This might show the well-known effect where a large negative potential shifts through zero (the iso-potential point) to a large positive potential as the pH is lowered. More interestingly, systematic measurements from varying the concentration of a small additive might show that the additive which was there for purposes not connected to the particle dispersion, might produce an unexpected and unwelcome large change of ζ value. It is precisely these unexpected consequences of well-intended additions of other components that the ζ rule is intended to reveal. If, under all reasonable formulation variations, your ζ is rock solid and if you have the right dielectric constant and a low level of salt additives, then you are likely to have few surprises in terms of particle stability. If you start to see significant variations then you can probably come up with a reasonable hypothesis of why additive X (e.g. via a minor ionic impurity) is affecting particle Y and do something rational about it. An especially common "minor ionic impurity" can be a surfactant. An anionic particle formulation showing robust DLVO stability can be instantly converted to a flocculated mess by the accidental addition of a small quantity of a cationic surfactant. The first time this happened to me, I didn't even know that my additive included a low level of surfactant, so I had no reason to suspect that (as it turned out) it contained a cationic version.

5.3 Depletion flocculation



One of the very real and depressing problems about adding polymers to provide steric stabilisation is that they can make matters worse if they break loose or are added in too large a quantity and start to float around in the solution. If the polymer "likes" the particle then it might find that it is touching and holding together two particles, causing "*bridging* flocculation". This is relatively obvious and not discussed further.

For *depletion* flocculation to take place, the condition is that the polymer does *not* like the particle. It is, ultimately, the absence of liking that causes the particles to self-associate.

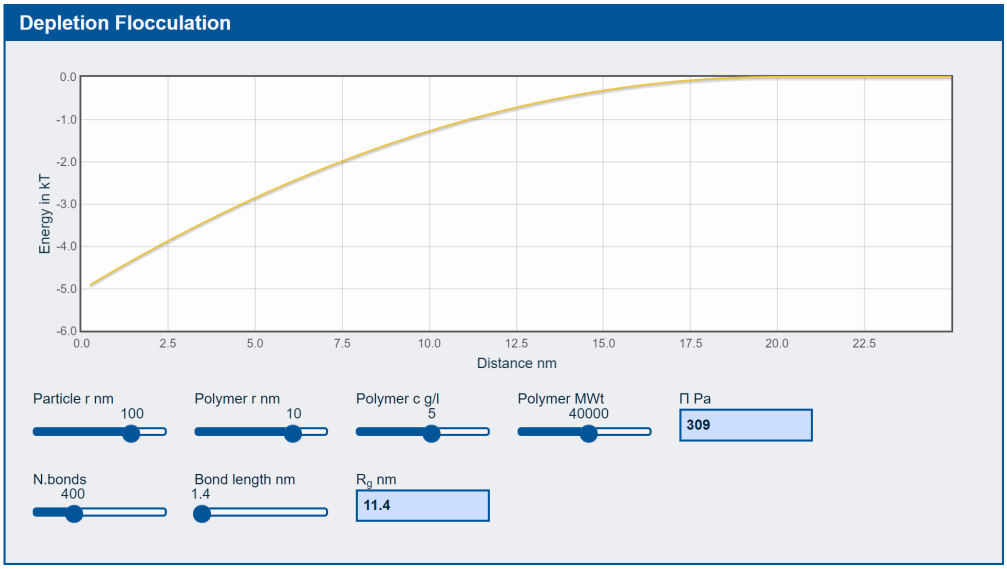


The image ignores any polymer that happens to be on the particle and focusses on polymer chains in solution. If the particles are close then the polymer cannot fit in between them, so there is a missing concentration of polymer and a net osmotic effect which causes the solvent to go from low to high concentration, drawing the particles together. One could equally describe the phenomenon in terms of excluded volume effects. But pursuing the osmotic pressure approach we can use the Asakura and Oosawa formula which is based on the radius r of the particle, the radius δ of the polymer chain, the centre-to-centre distance R between the particles and the polymer concentration and MWt from which the number density ρ (molecules/m^3) can be turned into the osmotic pressure Π via Π=ρkT. Asakura and Oosawa tell us that the attractive force G in values of kT is given by:

Equ. 5-5

$$G = -\Pi \frac{4}{3}\pi(r+\delta)^3\left[1-\frac{3R}{4(r+\delta)}+\frac{R^3}{16(r+\delta)^3}\right]$$

The somewhat inscrutable implications of this formula are best explored through the app which plots the particle-to-particle distance, h, (as in DLVO) instead of R, and automatically plots h between 0 and 2δ for which the formula is valid:



App 5-3 https://www.stevenabbott.co.uk/practical-solubility/depletion.php

The app calculates Π from the inputs. It also allows you to estimate the polymer radius δ from the standard approximation that $\delta^2=Nb^2/6$ when N is the number of bonds and b is the bond length. This should be the number of bonds of "statistical bond length" or "Kuhn length" rather than real polymer bonds. The calculation is added for your convenience and can be ignored if you have measurements of δ in solution.

5.4 Too much/little solubility

It is all very nice to describe dispersions in terms of idealised spheres of a fixed radius and polymers of a single MWt described by a single parameter such as χ. The real world is not like that and is invariably a set of compromises. Let us, for example, add a dispersing agent to some pigments to be used in an aqueous inkjet formulation. In the simple world of DLVO we might imagine that one end of the dispersing agent locks 100% onto the pigment and the other end is wonderfully happy in the water, so χ is low. In the real world, we know we have to add some co-solvent (often some sort of glycol) to, at the very least, stop the ink from drying out in the inkjet nozzle.

Suppose that glycol really likes the "pigment" end of the dispersing agent. Well, that end is no longer so well locked to the pigment and the dispersing agent might start to float off and, at the same time reduce the steric stabilisation and encourage depletion flocculation.

And suppose that the glycol makes the hydrophilic end of the dispersing agent less happy in the water - we can tip over the critical χ=0.5 and the pigment crashes out. It won't crash out in the bulk solution, that would be too easy. Instead it crashes out in the nozzle when some of the water has evaporated, giving the high glycol concentration which takes χ over 0.5. This means that you seem to have a perfectly stable dispersion yet one that clogs the nozzles.

This scenario is not something I made up. It is from the real world of inkjet and is described in a paper[44] given at the HSP50 conference. The image from one of the slides in the talk (created by Geert Deroover

44 Bart Wuytens of AgfaLabs, *Measurement of HSP as a commercial service*, found on https://www.hansen-solubility.com/conference/papers.php

and reproduced with his kind permission) describes all the trade-offs:

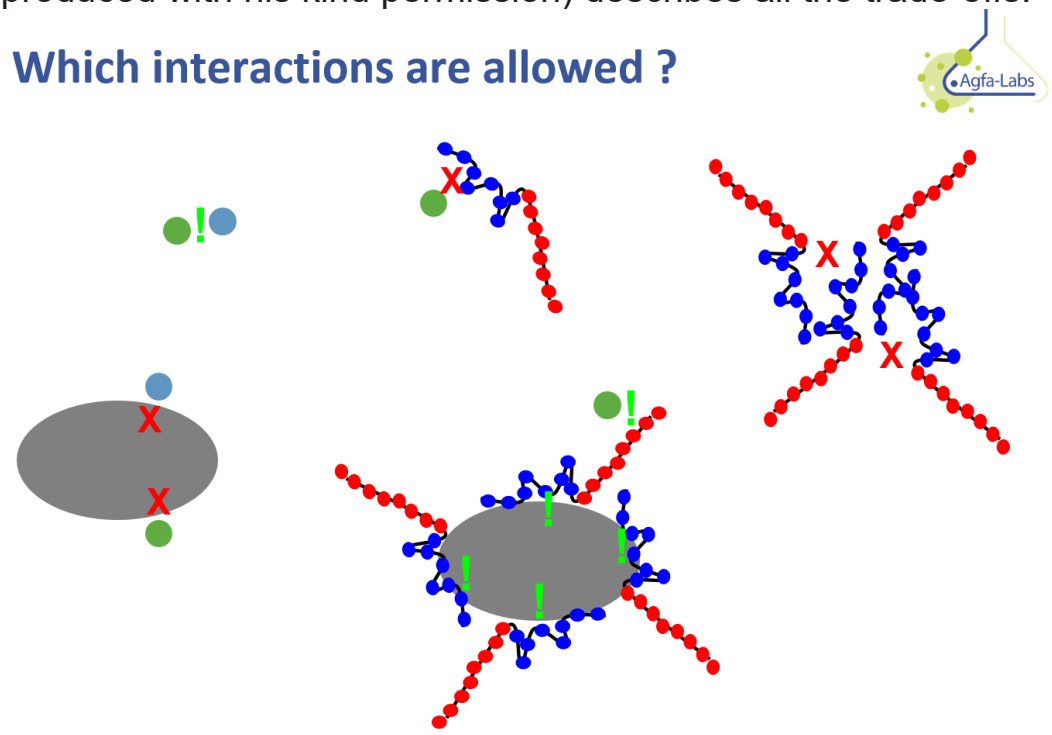
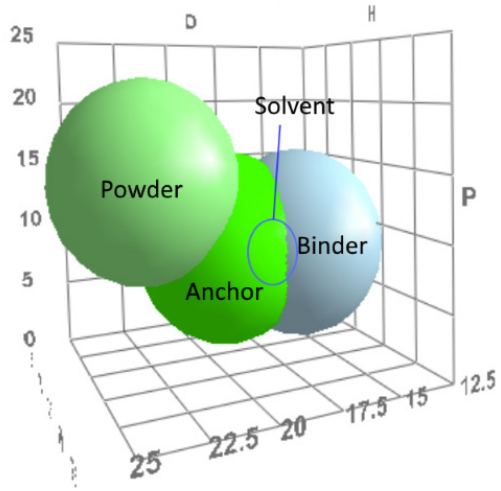


Figure 5-2 Desired (!) and undesired (x) interactions between the components of an inkjet ink

The x's mark interactions you don't want (e.g. water or glycol with the pigment) and the !'s mark those you do want (e.g. water must like the glycol). In the absence of any practical solubility theory, the formulator has to juggle things via "intuition".

With a theory such as HSP the formulator can get some key numbers and optimise the system accordingly. By knowing the HSP of the pigment, the respective ends of the dispersing agent and the various glycols that might be used, the impossibly large number of permutations of pigments, dispersants and glycols can be narrowed down to a few that meet the key criteria of not having too much solubility (pigment/dispersant) or too little solubility (hydrophilic_end/ glycol).



A very different application area, this time in a solvent dispersion, reaches the same conclusion - that you need to be clear what needs to remain attached to what in order to give the functionality your require. The example is a sophisticated ceramic of tantalum carbide. Just writing that sentence is a pleasing reminder that solubility science can appear in many unexpected places. Although readers can access the paper

describing the work[45], I have created a simplified version on the Hansen Solubility website (https://www.hansen-solubility.com/HSP-examples/ceramics.php) from which I have taken the diagram that summarises what is needed. The system has the ceramic pigment, a solvent blend (it is important to have different evaporation rates during the coating process), an "anchor" and a "binder". Too much or too little solubility will give an unstable system. The sweet spot is a solvent just between the anchor and the binder. The difference in coating performance between an optimum formulation in that sweet spot and one elsewhere within HSP space is dramatic.

The authors of the ceramics paper note that the idea of tuning the solvent to a "not too much, not too little" spot is not novel; Charles Hansen had pointed this out many decades ago when he was working in the paints industry. Such fundamental ideas stand the test of time.

5.5 Controlled non-solubility

When you really want dispersions to fall out of suspension, for example to purify dirty water, you usually want to add the smallest amount of the cheapest stuff to do the job in the fastest manner - without changing the problem from one of removing the particles to one of removing the stuff that removed the particles.

I was keen to write a deep, informative review of how solubility science can help us achieve this task. As I have only infrequently ever needed to know anything about the topic, and the little knowledge I had was decades old, I looked forward to learning some interesting new (to me) science.

Although I tried hard to find interesting things to say, I failed. I even failed to find a real distinction between the three key terms used to describe the phenomena: agglomeration, coagulation and flocculation. Authors seem to use the words indiscriminately, even those who seem to be suggesting that there is a difference.

Everyone states, as they have done for decades, the four principles:

1. Add enough of the right ions to swamp the Debye effect in DLVO. If you go to the DLVO app and enter the parameters shown earlier, at 0.05M salt the barrier is the desired 20kT. Now swap the salt from something like NaCl to $Al_2(SO_4)_3$ and change Z_1 to 3 and Z_2 to 2, the barrier decreases to 14kT. So if you want a general destabilisation at the lowest salt concentration, use polyvalent salts.
2. Add salts that specifically "poison" the charge layer. As explained in the Zeta section, no one really knows much about this complex layer, so what goes there and acts like a poison is unclear. If (as happens to be the general case)

45 Daisuke Nakamura, Keisuke Shigetoh, Akitoshi Suzumura, *Tantalum carbide coating via wet powder process: From slurry design to practical process tests,* J. European Ceramic Society, 37, 2017, 1175–1185

water contaminants are anions, then cations like Al can be good poisons. What is not clear is why the Al^{3+} ion is far less effective (in general) than the $Al(OH)^{2+}$ ion, i.e. why you need rather higher pHs so that the Al is tending towards the insoluble hydroxide. As mentioned above, an ionic surfactant with a charge opposite to the particle can be an effective poison.

3. Poison the charge layer with polyelectrolytes which will form bridges with other particles. This can be done with Al or Fe salts at the right pH, with natural sugar carboxylate electrolytes (e.g. alginates) or unnatural polyelectrolytes such as co-polyacrylic acid salts (anionic to remove cationic suspensions) or quaternary amines (cationic to remove anionic suspensions), or with large clay particles such as the bentonites.
4. Get your additives to start to precipitate out on their own. The "net" they produce can drag your particles out of suspension. Tuning the pH of Al or Fe salt solutions can get you into this mode.

Then however much theory is thrown at us, it all comes down to doing a bunch of experiments which place different amounts of different flocculants into jars containing the suspension and finding out which precipitates the most, in the shortest time, leaving the lowest level of flocculant contamination.

A further complication is that the amount of particulate in suspension affects the whole process in two ways. The first is the obvious one that more particles require more flocculant to, say, neutralise them. The second is that at higher concentrations, the flocculating system itself (rather like the "net" approach) can help further flocculation.

At the heart of the problem is the fact that just about anything that destabilises a particle can restabilise it at a higher concentration then destabilise it again at an even higher concentration. So add a small amount of a cation to an anionic particle and the particle becomes neutralised and unstable. Add a little more and the result is a stable cationic particle. Add a small amount of a polyelectrolyte and it messes up the zeta potential, add some more and you have excellent steric stabilization, add even more and you might get bridging flocculation. You might at some stage even get depletion flocculation to add to the confusion.

I regret that I have been unable to find anything more profound to say. Maybe the whole topic is so complex that it is not possible to provide practical models to help the formulator solve problems more rapidly.

5.6 Is that it?

I was rather surprised to come so quickly to the end of this chapter. After all, I'm really talking about colloid science and there are whole books on the topic.

One large area of colloidal science is that of surfactants and emulsions. I have discussed these at length in my Surfactant Science book so they are not included in this book, other than the use of surfactants as solubilizers.

Arguably, (there are colloid science books that make this argument) aqueous polymers such as proteins, starches and celluloses can be considered as colloids and, indeed, the word "colloid" comes from the study of a natural gum polymer. This makes my point that there really is no distinction between dispersions and solutions. Such systems are discussed in the aqueous solubility chapter.

So, yes, that is it for dispersions.

I was wrong!

When I came to the conclusion that there was nothing more to write about dispersions, that was in 2017. It wasn't until 2019 that I heard of Scheutjens-Fleer (or Self-Consistent Field) theory and it took me many months of 2020 to be able to work with it and add it to apps and to HSPiP.

For those who are interested to know more about particle/dispersion solubility and, especially SF/SCF theory, my free eBook Particle Science: Principles and Practice tells you more.

6 Solubilizers

We already know a lot about solubilizers because we have used KB theory to discuss in detail two types of these: small molecule hydrotropes (urea, nicotinamide etc.), and entrainers for scCO2. In those discussions it was necessary to use KB theory to dispel the myth that these worked by forming clusters, even though hydrotropes have a "minimum hydrotrope concentration" which is misleadingly like the Critical Micelle Concentration (CMC) for surfactant solubilizers.

6.1 Surfactant solubilizers

We can now look at the classic surfactant solubilizers safe in the knowledge that these do indeed work via clusters (i.e. the micelles) though interestingly we shall see that the undoubted efficacy of these micelles is bought at the cost of a relatively low per-molecule efficiency.

Before we start the discussion it is worth dispelling a key myth that is highly popular and self-evidently wrong: that micelles solubilize hydrophobic solutes by bringing them into the hydrophobic core of the micelle. Why is this (in general) self-evidently wrong? Because the micelle core is equivalent to something like dodecane, a rather poor solvent for most of the interesting hydrophobic solutes, pharmaceutical APIs, for which solubilization is especially necessary. The root cause of this myth is unthinking use of the word "hydrophobic". It is worth repeating the sentence that outraged me when I first read it: "X is hydrophobic because it is insoluble in water so we therefore dissolved it in ethanol". A "hydrophobic" molecule that is soluble in ethanol is unlikely to be soluble in dodecane and is, therefore, not going to enjoy the core of a surfactant micelle.

Given that many solutes will not be solubilized in the core, where are they solubilized? The answer often seems to be "in the area between the tail and the head" though, as we shall see, it can also be "in the head area". After all, for a popular solubilizer such as Tween 80, the head area is basically low MWt polyethylene glycol, a solvent that is frequently used in pharma and cosmetics.

With that background, let us use the approach developed by Bhat, Dar, Das & Rather at U. Kashmir[46]. The basic formula is well-known. Given the CMC of the surfactant and the Molar Solubilization Ratio, MSR (an input to the app but, obviously, a parameter to be determined from experimental data), the solubility S at concentration c of surfactant for a solute with (low) water solubility S_0 is given by:

---

46 Parvaiz Ahmad Bhat, Ghulam Mohammad Rather, and Aijaz Ahmad Dar, *Effect of Surfactant Mixing on Partitioning of Model Hydrophobic Drug, Naproxen, between Aqueous and Micellar Phases*, J. Phys. Chem. B 2009, 113, 997–1006

Equ. 6-1

$$S = S_0 + MSR\left(c - CMC\right)$$

It hardly needs an app for such a trivial formula, but first there is a complication to deal with and second there is a lot of insight to be gained.

The complication is that we tend to prefer to specify added surfactant in % weight, so we need the MWt of the surfactant as an input.

The insights arrive when we add extra information that should be known for any given system:

- the number n of (linear) carbons in the hydrocarbon tail from which the volume and surface area of the micelle core can be calculated via the Tanford equation giving the radius r in Å of the micelle: r=1.5+1.265(n-1);
- N, the micellar aggregation number, the number of surfactant chains in a micelle;
- the Volume and surface Area of the solute.

App 6-1 https://www.stevenabbott.co.uk/practical-solubility/surfactants.php

We can now do some extra calculations:

- The increase of solubility over the range of surfactant concentrations chosen by the app.
- The number of solute molecules per micelle. This is calculated as N.MSR
- How much of the core would be taken up by that number of solutes if they were all in the core. If this is more than, say, 50% we have some obvious problems.

- How much of the surface of the micelle would be taken up if the solute were entirely in the head region. Again we have problems if this is more than say, 75%
- K_m the micelle/water partition coefficient via K_m=MSR/[0.018 S_0(1+MSR)] where the 0.018 is the molar volume of water in l/mol.

Now we can see why we have an app.

The solutes/micelle are discussed below. The micelle-water partition coefficient, K_m gives some idea of how happy the solute is within the micelle, without saying where in the micelle it is.

The %Core and %Surface calculations allow a sanity check about the plausibility of assigning a solute to any given part of the micelle. Of course it assumes that the micellar aggregation number is not affected by the solute, but if we throw out that assumption then analysing the system becomes unmanageable, at least for a simple app.

From the U Kashmir papers some interesting numbers emerge. A typical genuinely hydrophobic solute such as naphthalene might give 16 molecules in a typical micelle, equivalent to 8% of the core volume - which is no problem. With erythromycin the solubilization added 136 molecules per micelle, far exceeding the core volume but being 80% of the surface volume, which is a more plausible place for such a molecule. With Naproxen/CTAB, the default values of the app, the core is ~10% and the surface ~40% so they are equally plausible places. Common sense suggests that Naproxen would most probably be in the intermediate zone as its solubility in alkanes is small (<1mM, similar to its water solubility) whereas its solubility in something like octanol is around 120mM.

This classic explanation of solubilization via micelles seems to concord with the general and erroneous mythology that clusters are needed for hydrotropy and solubilization in general. It is time, therefore, to see why there is no contradiction between the assertion that clusters always have a negative effect on solubilization and the fact that micelles can be superb solubilizers. Here is the key KB equation once again, with one significant difference:

$$\frac{\delta\mu_u}{\delta c_2} = -RT\,\frac{G_{u2} - G_{u1}}{N + c_2\left(G_{22} - G_{21}\right)}$$

Equ. 6-2

Because we are only interested in the situation above the CMC, we know that the concentration of individual surfactant molecules is negligible, so, following the analysis of Shimizu and Matubayasi[47] (but using my own, less precise,

47 Seishi Shimizu and Nobuyuki Matubayasi, *Hydrotropy: Monomer−Micelle Equilibrium and Minimum Hydrotrope Concentration*, J. Phys. Chem. B 2014, 118, 10515−10524

nomenclature for simplicity) we can just call 2 the micelle rather than the surfactant itself, and we immediately find a bonus effect. There is very little micelle-micelle "clustering" and, indeed, G_{22} is mostly the excluded volume of the micelle which is relatively large and negative. So unlike the hydrotrope case, the term involving c_2 is negative which helps solubilization. But note the N in the denominator, rather than the 1 in the classic equation. The micellar aggregation number significantly reduces the solubilization effect.

The G_{u2} term is interesting. Because 2 is large, the excluded volume effect should be large, but in this case there is no excluded volume because the solute is allowed inside the (virtual) molecule. On this basis, micelles are efficient solubilizers. But because the N appears in the denominator, the attraction of the solute for the micelle is strongly offset by the need to create the micelles in the first place.

One cannot argue with thermodynamics, but such an argument doesn't give a "feel" for what it means. So let us look at the same thing in a different way. For a lot of solutes and surfactants, the MSR is typically <0.2. If we take MSR=0.2 and a micellar aggregation number of 50 then solute/micelle is 10. So it takes 5 surfactant molecules to attract one solute molecule. If there were fewer surfactant molecules in each micelle (i.e. if clustering were less) then the per-molecule efficiency of a surfactant would be higher.

So what? Surfactant solubilizers work so who cares if their per-molecule efficiency isn't brilliant? Using the CTAB example from the app, the Max% slider has to be moved to 1.2% to get a 10x increase in solubility. That is "only" 33mM, which is impressive compared to the typical 2M required for classic hydrotropes. But a paper on nicotinamide cites a 100x increase in solubility at 2M, which is 240g/l, ~24%. CTAB would need to be 0.33M, i.e. 12% to match this, still an improvement but only a factor of 2 not a *massive* improvement.

The implications for rational solubilizer design will be discussed later in this chapter.

6.2 Solubilization: Solvents to pre-ouzo

A wonderful 2005 review of solubilization[48] by the Kunz team in Regensburg, paints a broad picture of what constitutes a solubilizer. It includes an example of acetone which, at its maximum measured molarity, represents an aqueous solution containing 70% acetone. As the title of the review implies, cosolvents are included, but to me, 70% acetone is beyond "cosolvents" and is much more just another example of a solvent blend that falls under conventional solubility theory.

---

48 P. Bauduin, A. Renoncourt, A. Kopf, D. Touraud, and W. Kunz, *Unified Concept of Solubilization in Water by Hydrotropes and Cosolvents*, Langmuir 2005, 21, 6769-6775

A 2016 review[49] by the Kunz team de-emphasises things like acetone but introduces newer ideas such as solvosurfactants and the fascinating pre-ouzo effect. Anyone interested in solubilization should read both reviews as they contain a lot of interesting examples. Here I want to discuss the principles raised from this wide landscape so that we can mutually learn to solubilize more effectively.

6.2.1 Solvo-surfactants

The first class to be discussed is the solvo-surfactants, a term most commonly associated with the Aubry team at Lille with a typical paper[50] being one that discusses the hydrotropy of the versatile glycerol monoethers. And the first obvious point to be made is that unlike classic hydrotropes like urea or nicotinamide, these are liquids, so might be thought of as solvents (which, of course, in the right circumstances they are). The second point is less obvious. By calling them solvo-surfactants their mode of operation is automatically assumed by many (but not, as we shall see, by the original authors) to be analogous to that of real surfactants. Indeed, the long-chain (and long glycerol) versions of these molecules, the polygycerol alkyl ethers are interesting surfactants. Because this name was coined before KB theory started to clarify that clustering is bad for classic hydrotropes there is a possibility that the name itself might be misleading in terms of mechanism.

The paper reveals that these molecules have a MAC, Minimum Aggregation Concentration (at which point there are a number of changes, including a plateau in surface tension), typically in the 200-300 mM range. Clearly we are very far from typical surfactants with values in the μM or low mM range. They also, just like classic hydrotropes, have an MHC, Minimum Hydrotrope Concentration, below which there is very little effect on solubility of their specific solute. As the authors point out, the fact that the MHC values are often significantly higher than the MAC "questions once again the link existing between the improved solubilization of the hydrotrope solutions and the supposed associative phenomena."

At the time of writing there are no published KB data on these systems. My guess, therefore, is that they work via the conventional mechanism which is statistical solute-hydrotrope interactions which then lead to solute-induced hydrotrope aggregation. And just as there are competing effects with classical hydrotropes, there are similar effects with the solvo-surfactants. As stated before, urea may well be less happy to associate with a given solute compared to nicotinamide, but it might win-out overall because it has zero self-clustering

---

49 Werner Kunz, Krister Holmberg, Thomas Zemb, *Hydrotropes*, Current Opinion in Colloid & Interface Science 22 (2016) 99–107

50 Laurianne Moity, Yan Shi, Valérie Molinier, Wissam Dayoub,Marc Lemaire, and Jean-Marie Aubry, *Hydrotropic Properties of Alkyl and Aryl Glycerol Monoethers*, J. Phys. Chem. B 2013, 117, 9262−9272

while nicotinamide significantly self-clusters. In KBI language, G_{u2} for urea might be lower than it is for nicotinamide, but G_{22} in the denominator is lower for urea, giving it a chance to win out. So it might be possible that these solvo-surfactants are less effective than they might otherwise be *because* they are somewhat surfactant-like! We simply don't know until we have the numbers.

6.2.2 Pre-ouzo systems

If you take a bottle of Greek ouzo or French pastis or absinthe and pour the clear liquid into water, it becomes cloudy as the alcohol-soluble anethole (the aniseed flavour, or thujone in the case of absinthe) becomes insoluble and phase separates into a light-scattering emulsion. As KB theory completely breaks down during phase separation there is nothing to be said here about such a phenomenon other than the fact that if a KBI starts changing very rapidly that is a good sign that phase separation is near.

The topic here is a zone where scientific light scattering measures (and a careful eye) detects some sort of unhappiness in the solution, i.e. plenty of self-clustering, but without phase separation. This is the "pre-ouzo" region. Our earlier discussion about fluctuation theory and scattering provides a hint that KB can apply nicely in this area.

The classic example is water-ethanol-octanol[51]. Water and ethanol are miscible as are ethanol and octanol, but we know that octanol is only slightly miscible with water. If various mixes of the three solvents are studied then there are large zones that are entirely uninteresting, such as zones with low octanol that are boring water-ethanol solutions, zones with low water that are boring ethanol-octanol solutions, and zones with low ethanol which are boring water-octanol 2-phase systems. Near the phase separation area where there is plenty of water and octanol but enough ethanol to change the situation, there is the pre-ouzo zone or, as it is sometimes called, a "surfactantless microemulsion".

So what? Conventional microemulsions, 20nm blobs of oil in water that form spontaneously, protected by a suitable surfactant shell are one way to get oil-loving solutes into an aqueous environment. Interestingly, and very confusingly, there are plenty of papers discussing "hydrotropes" when they are in fact discussing microemulsions. Microemulsions are discussed at length in my Surfactant Science eBook and are, in any case, rather too far from my definition of solubility to be discussed in this book. But because surfactants are often seen as "bad", a surfactantless microemulsion might well be a good thing and because the pre-ouzo effect is a key part of the Kunz team's view of hydrotropes it is important to point them out.

---

51 Thomas N. Zemb et al, *How to explain microemulsions formed by solvent mixtures without conventional surfactants*, PNAS, 113, 2016, 4260-4265

The solubility questions about the pre-ouzo state are partly about the mutual solubilities of the solvents involved. That is of no relevance to solubilizers in general. The point is that these pre-ouzo states can be solubilizers for other solutes, some specific examples of which will be discussed below.

The fundamental question raised by the Kunz team is whether the undoubted clustering (the team are expert users and interpreters of SAXS/SANS) in the pre-ouzo state is necessary for whatever solubilization takes place in that state.

The answer to that question[52] is rather complex and borders on the debate within the classical surfactant solubilizers about per-cluster efficiency (which might be reasonable) compared to the per-molecule efficiency which is probably not too high. As with the solvo-surfactants, at present we have only informed speculation and no KBIs with which to form firmer conclusions. My belief is that once we have KBIs for solvo-surfactants and for pre-ouzo clusters we will have a unified picture across the whole of solubilizer space which can then be used to help address the far more important question of how to design or select the optimal solubilizer for any given solute.

6.3 Designing the perfect solubilizer

These dry discussions on solubilizers are included in the book because solubilizers are increasingly going to be a key part of the formulator's job. As a specific example, let us discuss perfumes.

6.3.1 Fragrances

For centuries, fine fragrances have been developed using ethanol as the key solvent. It happens that the overwhelming majority of interesting fragrance molecules are scarcely soluble in water so ethanol is necessary to get a stable solution. As a bonus, when the fragrance is applied, the ethanol evaporates rapidly, leaving relatively little water to disturb the fragrance on, say, the skin.

Via an illogical application of "green" principles, "they" have decided that all that ethanol (which is, after all, classed as a Group 1 carcinogen by some organisations) is going to destroy the planet because it is a volatile organic compound. Therefore fragrances are having to shift to fully water-based formulations. The fact that the resources required to make this shift are substantial, whilst the environmental gains are minimal, entirely escapes the green movement. But fighting illogic is not a worthwhile activity for the large fragrance houses that provide most of the fragrance formulations around the world so they, and the large personal care companies, are all researching ways to solubilize fragrance molecules.

---

52 Seishi Shimizu and Nobuyuki Matubayasi *Unifying hydrotropy under Gibbs phase rule*, Phys. Chem. Chem. Phys., 2017, 10.1039/C7CP02132A

The Tchakalova team at Firmenich have published work on two different approaches. The first is via classical microemulsions and, as mentioned previously, this method and Tchakalova's approach to it are discussed in my Surfactant Science book so won't be discussed in this solubility book. The second approach is via solubilizers. These can be micellar surfactants, though consumers are more and more frightened of surfactants and, at the time of writing, one side of the Atlantic is especially frightened of ethoxylated surfactants and the other side is especially frightened of "sulfates" making it hard to find anything that works at all *at a price that consumers are willing to pay*. No perfumer is going to add classic hydrotopes like urea or nicotinamide. This leaves things like the solvo-surfactants and pre-ouzo style formulations as options.

A paper from Tchakalova[53] shows the rich behaviour of a pre-ouzo fragrance formulation. Instead of the "scientific" water-ethanol-octanol, the water-ethanol-fragrance system is used. As consumers do not like milky-looking fragrances, the formulation with the lowest acceptable level of ethanol is the pre-ouzo one, with clusters (from scattering experiments) in the 2nm range. The volatility profile of this type of formulation is fascinating because it helps explain an observation known for many years in perfumery: that the most volatile components are not necessarily the ones that evaporate fastest at first.

Although I cannot trace explicit publications on solvo-surfactants for fragrances, one can well imagine that there is plenty of activity in the area when so many of those molecules (such as the glycerol monoethers discussed earlier) are basically "green" molecules.

6.3.2 Pharma APIs

Moving to another topic, the classic use for solubilizers is within pharma for water-insoluble APIs (Active Pharmaceutical Ingredients). Urea, nicotinamide, sodium benzoate and so forth have been used along with the Tweens and newer "solubilizer polymers" such as BASF's Soluplus.

So finding a good solubilizer is a key formulation question which will continue to increase in importance as formulators are driven (rationally or not) to more aqueous formulations. To the best of my knowledge, the choice of the optimum solubilizer for any given system is largely based on trial and error, once obvious choices have been made such as whether or not a solid solubilizer can be used or whether a classic surfactant such as a Tween is acceptable. Given that the idea of hydrotropes has been around for more than a century the fact that we still have no objective method to select the best is unfortunate.

---

53 V. Tchakalova, Th. Zemb, D. Benczédi, *Evaporation triggered self-assembly in aqueous fragrance–ethanol mixtures and its impact on fragrance performance*, Colloids and Surfaces A: Physicochem. Eng. Aspects 460 (2014) 414–421

6.3.3 Non-correlations

The Kunz 2016 review makes the good point that at least we know what does *not* correlate with "hydrotrope efficiency". Before we examine the non-correlations we need to think through the definition of "hydrotrope efficiency". They define it as $\delta(\log(c_u)/\delta c_2$, i.e. the slope of the curve when adding more hydrotrope has a large effect. This definition is insightful but (as they admit) not the whole story. If, for example, urea has a high efficiency value, does that compensate for the fact that it happens only after adding (for no benefit at all) 0.4M urea to reach the minimum hydrotrope concentration? Is a surfactant with a smaller "efficiency" superior because it becomes effective at the CMC of 1mM? And if, as seems plausible, larger MWt hydrotropes with larger hydrophobic domains are more effective in terms of molar concentrations, is the advantage in molar terms outweighed by the disadvantage in % weight terms? In other words, should "hydrotrope efficiency" be $\delta(\log(w_u)/\delta w_2$?

A different measure of efficiency is the Setschenow coefficient, K_S commonly used in classical hydrotrope papers, where $\log(S/S_0)=K_S(c_H\text{-MHC})$. The bracketed right-hand term is the concentration of hydrotrope minus the Minimum Hydrotrope Concentration.

There is even a complication about enhancements of solubility via solubilizers. If you go to the surfactant solubilizer app and change the S_0 slider everything stays constant except for the partition coefficient K_m and the Increase in solubility. Going from S_0=0.5mM to S_0=5mM means that the same solubilizer goes from an increase of a factor of 10 down to an increase of a factor of 2. This tells us that it is meaningless to use enhancement factors to say that hydrotrope X is good for solute Y but bad for solute Z, unless the baseline solubilities of the two solutes are similar.

Even with these reservations about measuring the effectiveness of solubilizers, and the different measures of the effectiveness, the message from the Kunz paper is that there is no necessary correlation with:

- solubility of the solute in the neat hydrotrope (just think of surfactants and of urea, though the point applies to conventional liquid solubilizers);
- solubility of the hydrotrope in water;
- size of the hydrophobic portion of the hydrotrope.

The third non-correlation is my statement which somewhat contradicts what Kunz says, but I am deliberately including molecules like urea which can be excellent hydrotropes with no significant hydrophobic portion. Kunz goes on to say that molecules with large hydrophobic portions might fail to be good hydrotropes if you simply can't get enough of them into aqueous solution - in which case you have to add extra hydrotropes to help the hydrotrope. Citing

an example from Yalkowsky[54], the limited solubility of the otherwise excellent hydrotrope benzyl alcohol in water can be overcome by adding ethanol or propylene glycol. However, the marvellous 1000x increase in solubility in the case of rapamycin via this trick has to be put into context. The hydrotrope (benzyl alcohol) is present in the admirably small quantity of 1.5% but the "extra hydrotrope" is a 20:80 mix of ethanol:propylene glycol present as 50% of the whole solution. Given that rapamycin is even more soluble in pure ethanol, it would arguably be better to use that single solvent rather than a complex mixture. And although the final formulation gives a 1000x improvement, that is starting from a very low initial solubility, so it is only 10mg/ml (~1mM) in the enhanced aqueous system, when its solubility in pure benzyl alcohol is >400mg/ml.

It is good to have some non-correlations. It would be even better if we had some positive correlations. So far I know of only one paper, from the Barlow group at KCL[55], which attempts to produce such correlations. The paper is for a single solute and as a result a prediction was made about a potentially superior hydrotrope not included in the original test set. The prediction was vindicated - a major step forward.

The paper tests the increase of solubility of indomethacin in the presence of ten different hydrotropes. There were large differences in efficacy - from a 3x increase at 0.5M hydrotrope from sodium p-toluenesulfonate to a 470x increase from sodium nicotinate. The data confirmed the non-correlations with some of the suggested rules of thumb of the hydrotrope world. Instead, using 34 datapoints, an artificial neural network (ANN) was set to work to find out the key parameters for a successful hydrotrope for indomethacin. The ANN was then set to work on a different set of potential hydrotropes to find some winners and losers. One loser was caffeine which was predicted to have a near-zero effect on indomethacin solubility which proved to be correct. One chemically interesting winner was resorcinol, predicted to have a 2000x effect. To quote the authors: "On the basis of the ANN connection weight interrogation and the in-silico screening of the 16 test hydrotropes, the 'ideal' hydrotrope for indomethacin was identified as a low complexity compound having a pyridine ring as hydrophobe, with an alkyl-substituted amide moiety, and a low hydrogen bond acceptor count." Using a natural neural network, i.e. their own brains, they found a food-safe potentially effective hydrotrope: pydridoxine, vitamin B6. When tested experimentally, at 0.5M it provided a 727x solubility increase, far outclassing sodium nicotinate, the best from the original dataset.

---

54 Pahala Simamora, Joan M. Alvarez, Samuel H. Yalkowsky, *Solubilization of rapamycin*, International Journal of Pharmaceutics 213 (2001) 25–29

55 Safa A. Damiatia et al, *Application of machine learning in prediction of hydrotrope-enhanced solubilisation of indomethacin*, International Journal of Pharmaceutics 530 (2017) 99–106

The authors intend to extend the work to multiple solutes in the expectation of gaining a general purpose hydrotrope prediction capability.

6.3.4 No good ideas

So, after 100+ years of effort on solubilizers, apart from the Barlow paper we have not advanced much in terms of prediction/selection. This is partly because the situation is complex and partly because our ways of talking about them have been confused and confusing. With KB theory able, at last, to describe the whole range of solubilizers and with a de-emphasis on the idea of clustering, we can focus on the one thing that matters most which is a high G_{u2} at the lowest possible concentration of solubilizer, without a high G_{22} (or, possibly, with a high G_{21}) allowing for a shift in the debate about surfactant solubilizers by defining "2" as the micelle rather than the surfactant itself.

So, excluding micelles, if there is no obvious link between solute-solubilizer solubility, what gives large G_{u2} (and, possibly, G_{21}) values at the lowest-possible solubilizer concentrations, along with low G_{22} (and, possibly, G_{u1})?

I have been tempted to seek individual rules, for example, how to obtain a high G_{21} but two simple examples shows that it is not, in general, possible to isolate any specific G_{ij} for optimisation. One way to encourage a large G_{21} is to make 2 a "hard" ion so that it attracts a large hydration shell around it. This is, however, unlikely to help achieve a large G_{u2} if the solute is not a strong H-bond donor. A strongly hydrophobic 2 might encourage a large G_{u2} but might also create a large negative G_{21} which would be bad for solubilization.

At present, the sad conclusion is that as a community we have little idea of how to design an appropriate hydrotrope. As we will see in the discussion on the Hofmeister series for proteins (systematic solubility changes depending on specific ions) there was a similar situation. At the time that the mechanism behind the Hofmeister effects was unambiguously clarified via KB analysis there was great confusion about the various detailed mechanisms at work and, of course, KBI could only say which general effect (G_{u2}) was important, not the specific effect. Happily, concerted efforts have recently revealed many of those detailed mechanisms, complexities and trade-offs with the Hofmeister effects. With hindsight we can see that the detailed mechanisms accord exactly with the expectations from KB. With some similar efforts devoted to classical hydrotropes maybe we can also go beyond the KBI and gain a deeper understanding. The first step is to look in the right place. The considerable work that has gone in to looking at "water structure" and "clustering" we know, with the benefit of hindsight, could not reveal deep insights because the causes of hydrotropy are not to be found in those areas.

If there is something positive to be gained from all these negatives, it is the idea that we will all make much faster progress in finding the optimum solubilizer

now we are no longer looking in the wrong places for answers. Looking in the right places requires us to ensure we always get the relevant KBI during the experiments (not a great additional burden) and to choose sets of solutes and hydrotropes which feature a set of functionalities that could test some plausible hypotheses such as (see the section on scCO2 entrainers) donor/acceptor pairings. With modern high throughput techniques a concerted effort could tell us more within a year than we have learned in decades.

7 Aqueous solubility

We all know that water is a "special" liquid and that "no one fully understands water" and that it has water clusters which are helpful in understanding many aspects of water as a pure liquid. Unfortunately, as mentioned in previous chapters, the special nature of water can be invoked in a hand-waving way to "explain" all sorts of solubility phenomena. The aim of this chapter is to give the formulator as many tools as possible that are devoid of "special" character. And we can get off to a good start with a formula for predicting the solubility of organic compounds in water.

7.1 The Yalkowsky GSE

We have already encountered Yalkowsky in the discussions of ideal solubility. Here we use his wonderfully easy General Solubility Equation, GSE[56], based on the original ideal solubility equation (dependent on MPt, T_m), which is combined with one other easily measured or predicted parameter, $LogK_{OW}$, the octanol/water partition coefficient. The GSE predicts the solubility, S, of any uncharged organic compound in water:

$$\log(S) = 0.5 - 0.01(T_m - 25) - \log K_{OW}$$

Equ. 7-1

The factor of 0.01 in this equation differs from the 0.023 of the earlier equation because this equation is based on $\log_{10}$.If the MPt of the compound is less than 25 °C then it should be set to 25 to ensure that the term containing T_m becomes zero.

In a sense the equation *has* to be right because it just takes the ideal solubility, assumes good solubility in octanol and so, from $logK_{OW}$ the real solubility must emerge.

How good can such an equation be? The answer is that it has stood the test of time and compares favourably with much more complex formulae. The beauty is that $logK_{OW}$ is readily measured or estimated from the molecular structure and although it is hard to estimate MPts, they are easy to measure.

It would seem self-evident that any weak acids, bases or zwitterions would fail to be predicted accurately by the GSE. Yalkowsky developed correction factors[57] (they are simple, but I won't describe them here) for such weakly ionic organic compounds and found that using these correction factors helped only a little,

---

56 Yingqing Ran and Samuel H. Yalkowsky, *Prediction of Drug Solubility by the General Solubility Equation (GSE)*, J. Chem. Inf. Comput. Sci. 2001, 41, 354-357

57 Neera Jain, Gang Yang, Stephen G. Machatha,Samuel H. Yalkowsky, *Estimation of the aqueous solubility of weak electrolytes*, International Journal of Pharmaceutics 319 (2006) 169–171

with the average absolute error for such compounds going from 0.72 without to 0.67 with the corrections. The equivalent error for non-electrolytes is 0.38. What Yalkowsky concludes from this is that we can ignore whether the molecule is or is not a weak electrolyte and carry on using the simple equation, though being alert to the fact that weak electrolyte predictions will be less good.

What about the effect of electrolytes on the solubility? As far as I can tell, for organic molecules the effects are "small" compared to the errors of the estimation.

7.1.1 Poorly water-soluble drugs

There is a significant industry devoted to writing articles and books on improving the solubility of poorly water-soluble drugs. My take on the topic is that you cannot improve the solubility without changing the drug. That may be overly pessimistic but it saves a lot of effort trying to follow approaches that regularly fail. Here are some of the approaches. I am omitting ideas which effectively change the problem, for example, changing the form of the drug from a carboxylic acid to a salt. I regard that as creating a different molecule, and it is the solubility of a given molecule that concerns me here. Here are some popular approaches which simply do not work:

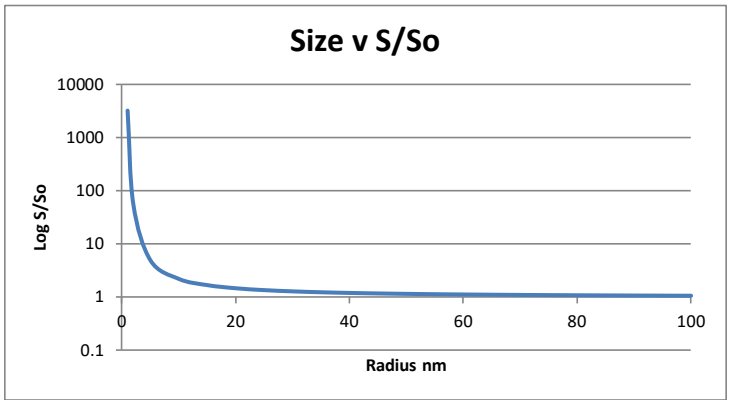


Figure 7-1 The theoretical increase in solubility with small particle size

1. Make the particle nanosized. The Ostwald-Freundlich formula tells us that for a particle with a surface energy of γ and radius of r you can increase the solubility to S from S_0 via $\ln(S/S_0)=2(\gamma/r)(MVol/RT)$. This seems too good to be true - you get more solubility just by making the particle smaller! And it is too good to be true. The effect only becomes significant for particles with a radius less than 5-10nm.It is seriously difficult to make particles that small and if you can manage to make them, the property which increases their solubility (a high surface pressure, γ/r) sets them up for Ostwald ripening whereby the smaller particles get smaller and the bigger particles get bigger. [An app for this is at https://www.stevenabbott.co.uk/practical-solubility/ostwald.php] The experimental evidence for such solubility effects is full of controversy (many very large enhancements are now known to be due to the inability

of the experimentalists to filter out undissolved nanocrystals) and even the best evidence I could find for enhancements in the 2-10x range struck me as rather indirect and not at all relevant to any real-world method for increasing solubility. This is simply not a practical solution to the problem.

2. Increase the kinetics of dissolution by making particles smaller. A tablet made from a single crystal of the drug will obviously dissolve much slower than one made from fine particles. But the literature shows that fine particles start to bring their own problems. If they clump together nicely via their large surface area then they can be *more* difficult to dissolve. In any case, most drugs are delivered in tablets with excipients so that a small amount of drug can be provided in a conveniently-sized package. The interactions of the drug particles with the excipients are likely to be far more important than merely having smaller particles.

3. Make the drug amorphous. The discussions on ideal solubility show that this should be a very easy win. And experiments have shown 5x or 10x increases in solubility in carefully-designed experiments. But you can't fight the laws of thermodynamics. In some of these experiments the solubility of the amorphous form is 5x larger after a few minutes then decreases to 2x larger as the thermodynamics kicks in and makes the drug crystallise out of the water. And good luck convincing yourself and the regulators that your amorphous form does not spontaneously crystallise during 2 years of storage on an over-warm shelf somewhere.

Now to four approaches that can work. Because you can't fight the laws of thermodynamics, these approaches work by changing the thermodynamics.

1. If the drug is a salt, change the counter-ion. This is an obvious and much-used tactic. The rules for the best counter-ion are probably impossible to determine because ultimately you are playing with crystalline packing forms which remain largely impervious to practical estimation.
2. Find a lower-melting polymorph of the drug and hope you can keep it stable indefinitely. By the ideal solubility law, this trick will work and, indeed, has been made to work for some drugs. The effects are not dramatic; from the ideal solubility app you can find that the effect on solubility of a, say, 10°C lowering of the MPt is modest. One problem is that your favourite polymorph today may become impossible to produce tomorrow. The stories are true of drugs existing as polymorph A for many years then being uncrystallizable into that form after the day someone appeared with polymorph B. It takes just one seed crystal in the wrong place ...[58]

---

58 I am always happy to find an excuse to cite the modern review built upon an equally delightful review about disappearing polymorphs, seed crystals etc. This article is Open Access. Dejan-Krešimir Bučar, Robert W Lancaster, and Joel Bernstein, *Disappearing Polymorphs Revisited*, Angew Chem Int Ed Engl. 2015, 54, 6972–6993

3. Co-crystallise the drug with some harmless other molecule so that the MPt of the co-crystal is very much lower than that of the drug, i.e. create a deep eutectic mixture (we discuss natural deep eutectic solvents, NADES, later). From the simple principles of ideal solubility, an increase in solubility is likely, especially if the enthalpy of fusion is also decreased sharply. There has been a boom in publications on co-crystal screening, i.e. rational approaches to finding the best co-crystal. Some papers have suggested that HSP might be useful but my own view is that this is unlikely to be too helpful. I think that matching MPts is probably an important requirement in order to get a deep eutectic, i.e. the MPts cannot be too far apart for a deep eutectic to be possible. And just as with polymer-polymer interactions my impression is that H-bond donor/acceptor interactions (which standard HSP cannot deal with) are more likely to have a significant impact. The COSMOtherm package includes a co-crystal screening option which takes a (typically 1:1) mix (as virtual solvents, i.e. with no knowledge of the crystal structures), and calculates the excess enthalpy as enthalpy of mix minus enthalpy of each of the individual compounds. Those pairs with a larger negative excess enthalpy are much more likely to show good co-crystal behaviour. Much of the excess enthalpy comes from hydrogen bonding (donor/acceptor interactions) but the results are best when all factors are included.
4. Solubilize the drug with any of the solubilizers or hydrotropes discussed earlier.

In all cases, *in vitro* experiments are guaranteed to show success. The only doubt is what happens *in vivo*. For the first trick, if a Rb salt gives wonderful solubility but the stomach is full of Na, will the Na salt crystallise out? For the second trick one can imagine cases where the stomach environment encourages conversion to the other polymorph, though this seems unlikely. For the third and fourth tricks, if the co-crystal or solubilizer molecules get whisked away by some biological process, will the original molecule then fall out of solution?

My reading of the literature is that the stomach is such a dynamic environment that any trick which affects the *kinetics* of dissolution is likely to give some benefit to the amount of drug that ends up in the bloodstream. This means that if it is not possible to find a better salt, polymorph, co-crystal or solubilizer then help should be found by focussing on how the drug particles are dispersed within the excipients of the tablet, and how the excipients interact with the contents of the stomach. Not much to do with solubility but still valuable.

7.2 Hofmeister horrors

Everyone agrees that in aqueous solutions that involve large ionic species (proteins, DNA, nanoparticles, colloids, bubbles) you can often find that stability (or conformation etc.) changes significantly with a change in counter-ion, in a

relatively constant order, with larger changes from the anion series than the cation series:

$SO_4^{2-} > HPO_4^{2-} > F^- > CH_3COO^- > Cl^- > Br^- > NO_3^- > I^- > ClO_4^- > SCN^-$

$(CH_3)_4N^+ > NH_4^+ > Rb^+ > K^+ > Na^+ > Li^+ > Mg^{2+} > Ca^{2+} > Guanidinium^+$

When Hofmeister first discovered this series and its links to effects (e.g. activity coefficients) of these ions in water it seemed rather clear that the effect was due to the "water withdrawing power of salts". That quoted phrase is from Hofmeister but is found in the masterly review of Hofmeister effects by Kunz[59] that summarises a set of papers from a colloquium on that topic. Unfortunately, Hofmeister himself could not work out how "water withdrawing power" caused the effects. And more than a century later the situation remained highly confused.

Reading the colloquium papers is sobering. Fine minds who have thought deeply about what is going on emphasise that the "explanations" that are generally advanced for Hofmeister effects simply don't work. Readers of this book who have looked at my take on DLVO and zeta potential might have noticed that I treated both effects with an absence of theoretical rigour. That is correct, but it happens to be entirely justified. Most of the assumptions behind DLVO do not withstand rigorous examination when examined through the lens of Hofmeister effects. As the Hofmeister review papers emphasise, things like Debye lengths or the neat separation between van der Waals and charged effects simply cannot withstand detailed scrutiny. A specific quote from the review paper makes this clear. "[... after estimating the Debye force and the van der Waals force] (w)hat remains after subtraction is called a ‘hydration force’. However if we subtract the predictions of two incorrect theories from a perfectly good measured force curve, we have in fact not a hydration force, but rather, non-sense."

The list of "large ionic species" mentioned above included bubbles. If you create a stream of bubbles in pure water they quickly join together and collapse. The same stream of bubbles in water containing salts become increasingly robust to collapse, following the Hofmeister series. This is a system far simpler than a protein or a colloid particle, yet the reasons for this clear Hofmeister effect are still largely unknown.

The expert view (including that of Hofmeister himself), therefore, was that Hofmeister effects are deeply hard to understand. If, however, you read *non*-expert views on Hofmeister, a very different picture emerges. Authors speak with confidence about "chaotropic" and "kosmotropic" ions and the explanations

---

59 W. Kunz, P. Lo Nostro, B.W. Ninham, *The present state of affairs with Hofmeister effects*, Current Opinion in Colloid and Interface Science 9 (2004) 1–18

invariably include water structure, clusters and H-bonds. These papers always made me feel stupid. I have been reading about chaotropes (these are said to disrupt H-bond networks in water) and kosmotropes (these are said to make the networks stronger) for decades and not once did the explanation ever make sense to me. I could never pin down how the breaking or making of H-bonds in one part of the system explained the change in another part of the system. And I suspected that the authors didn't know either. Discussions invariably led to entropic effects that, perhaps, almost but not quite cancelled out enthalpic effects, the net effect of which was a greater or lesser stability of the particles in question. I even get confused by the terminology used to describe the effects. The conclusion of the otherwise excellent Okur et al paper described shortly says "*... the more strongly attracted [the] anion is to a protein in solution, the less efficient it is in its salting out and vice versa*". "Efficient in salting out" means "Efficient at making something less stable in solution". This happens, also, to make it "Efficient at avoiding denaturation", which often gets translated into "Efficient at making the protein more stable". To use the phrase "more stable" for something which causes the protein to fall out of solution seems to me to be unhelpful language. It has taken me some time to get into the habit of translating sentences like the one from Okur et al into "... the more strongly attracted [the] anion is to a protein in solution, the more efficient it is in solubilizing/denaturing/unfolding and vice versa". It is this nightmare of non-explanations and confusing language that I call Hofmeister horrors.

If you think I am being harsh in talking about Hofmeister horrors, a quote from one of the symposium's papers[60] examining the bubble effect may make my views seem less unreasonable: "Our understanding of ion specificity has barely progressed from the empirical work of Hofmeister more than 100 years ago, with the terms 'ion size' and 'polarisability' giving names to our ignorance, in the absence of an ion-specific theory."

Fortunately, science advances, and since that colloquium some useful insights into chao/kosmotrope ideas came in a later review[61] on Hofmeister effects which describes very clearly why we still do not have a clear view of what causes these effects. The review explicitly mentions that ideas of how chao/kosmotropes affect long-range water structure are dead and that instead the terms simply mean "soft" ions that have a low degree of hydration (the SCN^- end of the anion scale and the $(CH_3)_4N^+$ end of the cation series) and the "hard" ions with a high degree of hydration (SO_4^{2-} and Ca^{2+} respectively). Some of the anomalies of Hofmeister series can be explained by Collins' Law of Matching Water Affinities[62]

---

60 Vincent S.J. Craig, *Bubble coalescence and specific-ion effects*, Current Opinion in Colloid & Interface Science 9 (2004) 178–184

61 Andrea Salis and Barry W. Ninham, *Models and mechanisms of Hofmeister effects in electrolyte solutions, and colloid and protein systems revisited*, Chem. Soc. Rev., 2014, 43, 7358--7377

62 Kim D. Collins, *Ions from the Hofmeister series and osmolytes: effects on proteins in solution and in the crystallization process*, Methods 34 (2004) 300–311

which says that hard-hard and soft-soft systems are different from hard-soft and soft-hard. This "law" turns out to be rather insightful. A very simple example is the dependence of aqueous viscosities of salts. If the ions interact strongly with the water then there is a large "Jones-Dole B coefficient" (which is a measure of the non-linear viscosity dependence), if they have only a weak interaction the effect of the salt is close to ideal behaviour and the B coefficient is small. Collins was able to show that salts like CsF or LiI had large B coefficients because the unmatched "hard" ion (F^- and Li^+ respectively) strongly influences the viscosity. But CsI (soft-soft) and LiF (hard-hard) have almost no effect on the viscosity. The two soft ions have little effect on the water and the effects of the two hard ions somehow cancel each other.

Then in 2017 the picture has become even clearer. A review[63] (which includes the satisfyingly clear statement: "Therefore, the whole concept of “kosmotropes” and “chaotropes” may need to be set aside.") shows that the specific ion effects on proteins can all be explained by specific sets of interactions with key functionalities on the proteins. Anions interact (of course) with cationic sites on the protein and this follows the obvious order of harder anions having stronger interactions. But the soft anions have significant interactions with the backbone peptide bond (with the C=O and NH groups) so for proteins with a low number of cationic sites on the protein, the Hofmeister effect is reversed, with the softer anions solubilising the protein. Cations interact as expected with the carboxylate groups on proteins, so harder cations are more solubilising. Because there is no competing interaction with the protein backbone, the Hofmeister effect is more straightforward for cations.

Note that there is a rather startling conclusion. The explanation of the Hofmeister effects of ions on proteins is via ions interacting with specific groups (some ionic, some not) within the proteins. This seems an absurdly simple thing to say. But one theme of this book is that the solubility community has invested vast efforts in complicated explanations when the simple explanations really work rather well.

7.2.1 The Potential of Mean Force

I need to introduce a new term at this point, the "potential of mean force". I have spent most of my scientific life unaware of it and regret my ignorance. To explain what it means, and why it is important it is necessary to look at something more familiar. We are all rather comfortable with the sorts of ideas shown in the DLVO app. We have 3 potential forces, each described in a way that makes plenty of sense (even if the details escape us) and these change rationally with distance. To know what happens at any distance we just sum the forces. Easy!

---

63 Halil I. Okur, Jana Hladílková, Kelvin B. Rembert,Younhee Cho, Jan Heyda, Joachim Dzubiella, Paul S. Cremer, and Pavel Jungwirth, *Beyond the Hofmeister Series: Ion-Specific Effects on Proteins and Their Biological Functions*, J. Phys. Chem. B 2017, 121, 1997−2014

Let us now imagine some real ions around real particles as they come together. Those ions happen to be parts of molecules and those molecules have finite sizes, and if molecule A is in a certain position that constrains where molecule B might be. There is zero chance that as those real systems come together the nicely defined van der Waals and Debye potentials will carry on operating as if we had a continuous medium. Instead, we need a potential that describes the messy reality of all those molecules. This is going to be some sort of mean potential that is certainly *not* going to have the shape of the idealised potentials. We just need to remind ourselves of the simplest radial distribution function of billiard balls around other billiard balls (or Lennard-Jonesium for those who do basic molecular dynamics). Here is the simple RDF from the KB chapter.

There is absolutely nothing idealised about this distribution even in such a simple system.

The superimposed illustration from the cover describes those effects more vividly:

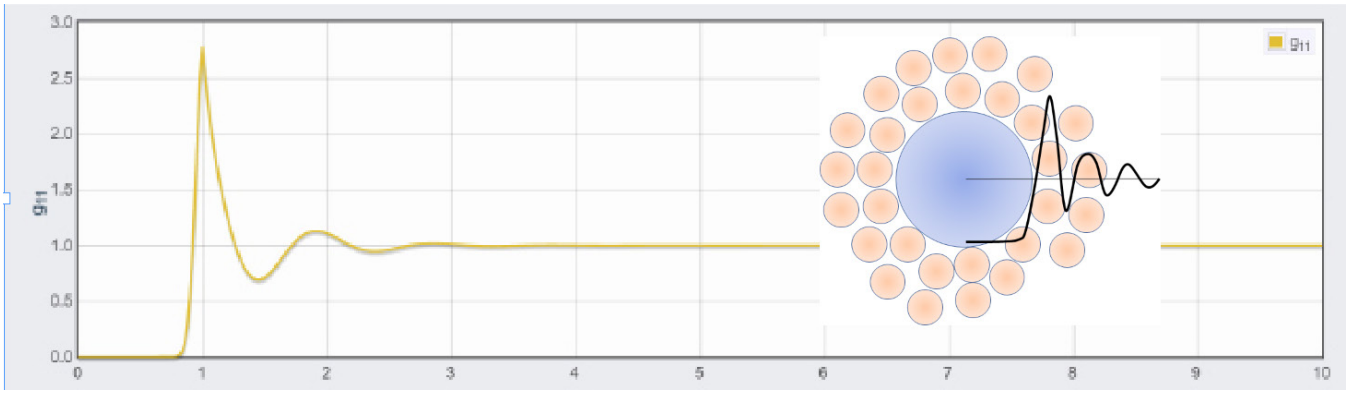


Figure 7-2 The RDF as a reminder that the potential of mean force is more realistic than mean field. Superimposed is a graphic showing where the RDF comes from

Up to the distance of 1 (idealised atom) there are no other atoms. Then there is a nice cluster of atoms around the central atom, then a dip because further atoms are excluded, then a small rise in a second shell and so forth.

The interactions between molecules in such a discontinuous system certainly can be described by a potential. But it is a *potential of mean force*, a statistical average.

This gives us a useful clue to the literature. If we have a complex problem, especially with molecules/entities of very different sizes, and it is described via potential of mean force, the chance is that the description is attempting to take into account the reality of the situation. If it is merely described via nice potentials, the chances are that the tricky reality is not being included. The Hofmeister reviews I have quoted are well-aware of mean force issues. One specific point is that DLVO is fine for "long" distances at low ionic concentrations

because the Debye forces are large enough for the mean force potential to be equivalent to the idealised potential. Indeed, ion-specific effects are not found at "low" concentrations of ions for "standard" colloidal particles. The problems arise when concentrations are higher (>0.1M) or (as with a protein or [discussed later] an ionomer/polyelectrolyte) there are strong local concentrations of ions. In both cases the forces act over a shorter range, so potential of mean force becomes relevant. The work of Christenson (" DLVO-theory is completely inadequate") mentioned in the main discussion on DLVO showed this point clearly. Careful measurements of the forces between two mica plates showed strong oscillations at distances <5nm, looking similar to what would be expected from a typical RDF.

In general those who espouse chaotropes and kosmotropes are unaware that the granularity of real molecules can have a large effect on the issues at hand. This is especially true about excluded volume effects. The zero value of the RDF "inside" an atom or molecule has a profound effect on the potential of mean force, especially when the excluded volume is that of a large protein or colloid particle. This is another way of saying that we should abandon the hand-waving terms that have achieved surprisingly little and start to use an assumption-free thermodynamics.

7.2.2 The KB view of Hofmeister - and more

An assumption-free and rather clear way of thinking about effects involving potentials of mean force is the KB approach. While it does not address any of the issues of van der Waals or charged interactions it does provide a clear framework for discussions about what effects are, or are not, significant. The fact that right from the start it shows that "water structure" (as generally understood) is completely irrelevant means that the horrors of "chaotrope/kosmotrope" can be swept away. By knowing (as we will see) the key effects, theoreticians can focus their energies on those effects.

The relevant theory[64] is simple to write and the conclusions from it are equally simple. The version I show here misses out some interesting details and uses a slightly different language for clarity. For conformity with the terminology used elsewhere in this book I will use Shimizu and Matubayasi's standard nomenclature rather than the one used in their paper. So 1=water, 2=Hofmeister salt, u=protein (or any other large particle). Note that for convenience (with no loss of rigour) the salt is considered as one species.

From (say) osmotic pressure experiments with changing molar concentrations of the protein, it is simple to measure $\delta c_1/\delta c_u$. It turns out that from these measurements we can calculate G_{u1}, the protein-water KBI. By measuring the

64 Seishi Shimizu, William M. McLaren and Nobuyuki Matubayasi, *The Hofmeister series and protein-salt interactions*, J. Chem. Physics, 124, 234905 (1-4) 2006

densities of the protein solutions we can measure the protein's MVol, V_u and because this is related to G_{u1} (known) and G_{u2} (unknown) via V_1 and V_2 (both measurable or readily estimated) we have G_{u2} our by now familiar measure of how the solute (protein) and other component (Hofmeister ions) interact.

Finally we can work out the key issue, the change of chemical potential of the protein with respect to Hofmeister ions. The paper, for historical reasons, uses "preferential hydration parameter" $d\mu_u/d\mu_1$, here I use the more intuitive "preferential interaction parameter" $d\mu_u/d\mu_2$ which depends on the KBI via the simple equation:

$$\frac{d\mu_u}{d\mu_2} = -c_2\left(G_{u2} - G_{u1}\right)$$

Equ. 7-2

The interpretation is very straightforward. If G_{u2} is large, i.e. if the ions love to be around the protein, and G_{u1} is modest (i.e. water is fairly happy around the protein) then the protein's chemical potential gets lower (because of the $-c_2$ in the equation) the more ions are added, meaning that the protein is more thermodynamically stable. As it happens, both G_{u2} and G_{u1} are negative because of the large excluded volume effect of the protein. So a "large" G_{u2} means "less negative than it might have been".

We now, as discussed earlier, have to be very careful with our language. Most of the literature on protein stabilities says that for greater stability you want ions that do *not* interact with the protein. The difference in language depends on your point of view. I am staying with a rational *solubility* point of view. Ions that interact strongly with the protein make it more soluble so, in my language, more thermodynamically stable. From the point of view of a functional protein, these good ions are bad because they tend to denature (unfold) the protein by welcoming it even further into the aqueous environment. So "salting out" is seen as "making the protein more stable" because although it falls out of solution it is functionally intact (folded). "Salting in" often denatures (unfolds) the protein so it is said to be "less stable". You can choose whichever convention pleases you. I am going to stick with the solubility convention which says that a large G_{u2} makes the protein more soluble, with the possibility of unfolding it.

When the data are examined, G_{u1} shows no significant dependence on the different ions so the big effect on the protein is via protein-ion interactions via G_{u2}.

This rather straightforward conclusion tells us two important things about the Hofmeister effects on proteins (and presumably all such systems). The added salt is having no significant effect on the protein-water interactions. And water structure, G_{11} is, as always, of no significance as it does not appear in the

rigorous equation. The Hofmeister effect is, therefore, simply one of relative attraction of the ions to the protein.

Because the argument says nothing about ions it can be applied to "denaturants" such as urea and "stabilisers" like trehalose or PEG. The KB analysis shows exactly the same. Urea has a large G_{u2} so urea is happy to cluster around the protein, lowering the chemical potential of the protein, making the protein system more thermodynamically stable, though the protein itself is denatured by unfolding. With large, neutral, molecules such as trehalose and PEG there is little chance for G_{u2} to be large even if the molecules happened to rather like the protein. The reason is simple - the excluded volume effect guarantees that G_{u2} gets off to a very poor start in the KBI, so only exceptional mutual affinity would create an overall larger G_{u2}.

KB's indifference to whether we are discussing ions or neutral molecules is significant. We don't have to think about Hofmeister as one phenomenon and neutral molecules as another. They are one phenomenon, a fact that is very liberating in terms of looking for deeper explanations. Although it is good to look for specific interactions, we can start with a null hypothesis of no interactions to see what might happen. A large sugar molecule might be equally happy interacting with water and protein so there is no overall "specific interaction" effect, yet the excluded volume effect via its size can drive a significant shift in the protein's conformation. If a specific sugar shows an effect away from the null hypothesis then it is time to invoke specific protein-sugar interactions.

Is that it? Yes and no. Yes, because it allows those who wish to unravel the complexities of Hofmeister effects to focus on the only thing that really matters which is the effects on G_{u2}. No, because KBI entirely mask the reasons for a given value. Shimizu is very clear about this and also says that the interactions of the ions with the protein are "water-mediated". This admits, of course, that water-ion-protein interactions are key, precisely because KBI is a theory of potential of mean force where all the players have a role. So the KB approach does not itself solve the problem. Yet the complete absence of invocation of chao/kosmotropy and the absence of the word "hydrophobic" in the KB explanation means that the debate can shift to a new level of clarity.

As it happens, the Shimizu paper came out at about the same time as Kunz's Hofmeister review. It was clear to no one at that time that the KB view and the "specific ion effect" view would converge so nicely. KB says "it is protein-ion interactions that matter" and the later Okur et al paper says the same thing, whilst providing the detail of those interactions.

This illustrates a guiding principle. KB experiments can be done with rudimentary equipment (osmometers, densitomers) and calculations can be done even with pencil and paper. Out of this simplicity comes, generally, clarity as to root cause. Experimentalists with access to more complex equipment

and computer models can then focus their efforts on disentangling the specific reasons behind (say) a large G_{u2}. This is a far more productive approach then trying to prove vague hypotheses with experiments whose relations to the vague hypotheses is unknown.

Shimizu and colleagues have not stopped here. Phenomena such as denaturation or gelling imply a change in G_{ij} values and in their gastrophysics[65] paper that discusses the gelation of tofu, the story is dominated not by G_{u2} but by ΔG_{u2} - the difference of G_{u2} between states. The fact that it does not depend on ΔG_{u1} (i.e. on changes in hydration) again allows the discussions to focus on how ions (or, indeed, urea, sugars, PEG etc.) interact with the protein in the two states. For most of the larger molecules the answer is easy - the excluded volume effects so dominate (the null hypothesis) that it takes something exceptional to produce a positive interaction that affects the situation. One example is that the precise position of the -OH groups on various similar sugars has a significant effect on the ability to interact with two separate protein molecules, i.e. to act as a sort of bridging flocculant. So although, in general, KB analysis does not point to specific mechanisms, the fact that it can disambiguate mechanisms and give general clarity can sometimes enable a root cause to be identified.

Does any of this matter? A 2017 paper[66] raises the level of interest from nice discussions of proteins in test tubes to fundamental questions of biological processes. We all know that ATP (adenosine triphosphate) is vital for providing biological energy. It can do this at μM concentrations. Yet it is present in cells at mM concentrations. The paper reveals that ATP acts as a classic hydrotrope (comparable in behaviour to sodium xylene sulfonate) but at concentrations 10x lower. The authors speculate that ATP might be a universal solubilizer for proteins, required from the earliest days so that concentrated protein solutions could function in cell-like environments. They then show that ATP helps keep amyloid and prion proteins from phase separating. The implications for Alzheimer's and BSE diseases are clear. What interesting research prospects this paper has opened up!

7.2.3 The KB view of biologics (biopharmaceuticals)

The era of small-molecular pharmaceuticals seems to be slowly coming to an end with the rise of the biologics, which are proteins or other complex biopharmaceutical products. They aren't new; vaccines are biologics. But the newer biologics are proving to pose many problems in terms of productionising them and in registering them as safe and efficacious and, if you want to produce

---

65 Seishi Shimizu, Richard Stenner, Nobuyuki Matubayasi, *Gastrophysics: Statistical thermodynamics of biomolecular denaturation and gelation from the Kirkwood-Buff theory towards the understanding of tofu*, Food Hydrocolloids 62 (2017) 128-139

66 Avinash Patel, Liliana Malinovska, Shambaditya Saha,Jie Wang,Simon Alberti, Yamuna Krishnan,Anthony A. Hyman, *ATP as a biological hydrotrope*, Science 356, 753–756 (2017)

a generic version when the original goes off-patent, in terms of proving that your biologic is the same as the approved version.

A key issue is that the task of productionisation requires the biologic to be shaken, sheared, filtered, cooled, heated, mixed with air or solvents - just about anything that might disrupt a delicate protein structure. Once the product is safely in its final form, it has to survive being transported and sitting around on pharmacists' shelves before use.

Faced with this challenge of stability throughout processes and in storage, the formulators have to try a host of formulation additives: surfactants, sugars, amino acids, salts, osmolytes ...

This is a deeply serious, high-stakes activity, so what solubility science is used to optimize the process?

The situation is complicated because there are at least two competing effects. The first is the possibility of specific segmental interactions that can cause the protein to irreversibly lose its native conformation. If one sequence of peptides can create a nice β-sheet with another sequence, and if the protein has a chance to (temporarily) unfold and for the two sequences to come together into the irreversible form, then the protein is effectively lost. There is immense computational effort going in to identifying potential complementary pairs to understand the risks of this happening. If the effect goes to the extreme then the protein ends up as insoluble fibres; this "fibrillation" is characteristic of the amyloid structures in Alzheimers.

The other effect is the conventional issue of unfolding (denaturing) or of agglomeration. Unfolding would be easy to inhibit using, for example, "excluded volume" additives such as sugars, if it were not for the problem that such additives also can encourage the proteins to agglomerate. For those proteins that have a tendency to fibrillate, any formulation would have to push as much as possible towards reduction of unfolding without tipping over into agglomeration.

At the hand-waving level, all those in the field are aware of what needs to be done. The problem is the lack of rational tools to know how best to steer the proteins only in the specific direction required. Reading papers on the topic quickly becomes tedious. Some of them are biased accounts of the magic ingredient (say an alanine/glutamate mix) that works for the one or two proteins discussed in the paper and which, you can be sure, will not be the ingredient of choice for most others. Others show that A is OK for this aspect, B is OK for that aspect and surprisingly an A+B combination shows some unexpected effect. To label these papers as stamp collecting is a little unfair given that the problem is so hard, though effectively that is about as good as it gets. As this is a hugely

important issue there is a meta-industry of trying to analyse all the data to spot patterns.

My take on it is that there is too much gathering of incomplete data with insufficient analysis, so data mining will remain disappointing. It reminds me of the (far less challenging) situation with scCO2. Patterns only emerged when a "good enough" analytical tool (ultimately rooted in KB) was applied to the 100+ papers that had plenty of data with no common theoretical approach to make sense of it.

In the biologics world there is some recognition of KB. In particular, the G_{uu} value for protein-protein interactions is recognised as the assumption-free way to analyse light-scattering data. Simpler techniques such as B_{uu}, the second virial coefficient, contained too many assumptions so it was hard to unify the data. If there are data on G_{2u} and G_{22} values then I have failed to find them, and without such data the field is working semi-blind.

Given that getting hold of enough pure protein to do these experiments is hard enough, it sounds mad to suggest that everyone should rush out and measure all the relevant G_{ij} values. That is not what I am suggesting. Instead, the suggestion is that any time these experiments are carried out, the (usually modest) extra amount of work needed to extract the G_{ij} values should be a routine part of the process.

It comes down to a choice. Carry on stamp collecting with hand waving theories. Or start with an assumption-free theory that has proven to be insightful everywhere it is used, and do the extra bit of work (on top of the huge work needed to get enough protein to do the experiments) to get all relevant values. The realisation that G_{uu} is a more reliable guide than the alternatives is a good start.

I know which approach I would choose if I had to bring biologics into production and sale. And if I were the FDA, I would be delighted if the biologics world could provide the relevant G_{ij} values as part of their reasoning of why adding 1% of X was required to bring the biologic through the production process and to ensure long-term storage stability.

7.2.4 The KB view of benzene salting out

If you add the chlorides of Li, Na, K, Rb, Cs to a solution of benzene in water, the more salt you add, the less soluble the benzene becomes. This is classic "salting out" and mostly conforms to Hofmeister. Explanations for this effect cover the usual hand-waving ideas of "water structure" and "hydrophobic effect", each of which says precisely nothing about what the salts are actually doing. In this particular case, it happens that Li is far less effective than it should be,

so the salting out trend is $Na^+ > K^+ > Rb^+ > Li^+ > Cs^+$. As you can imagine, the water-structure explainers have an even tougher time.

A rather clear KB explanation of the effects can be found in a very readable PhD thesis[67] covering many interesting KB issues. There is a whole chapter devoted to the benzene problem and it methodically shows how simple the explanation is.

The problem is studied via molecular dynamics. This can often be a recipe for the generation of artefacts, especially given the notorious difficulty of going from RDF to KBI. Unfortunately, for the key KBI that matter, the problems of simulating ions accurately were not overcome so we only have some general pointers towards a positive explanation. However, the KBI for other aspects of the problem are good enough to disprove alternative explanations.

With water=1, salt=2 and benzene=u, naturally enough G_{11} was irrelevant and, changing from the thesis' nomenclature to the one used throughout this book, the change of solubility of benzene depends on $(G_{u2}- G_{u1})$ divided by $(1+c_2(G_{22}- G_{21}))$. It turns out that the salt-salt (G_{22}) and salt-water (G_{21}) terms follow Hofmeister fairly well (so cannot explain the oddity of Li) and the overall effect of the denominator is small, with $(G_{22}- G_{21})$ showing only a modest change, further reduced by the small value of c_2. So yet again the effect depends on $(G_{u2}- G_{u1})$, the relative interaction of the benzene with the salt and the water. G_{u1} varies only slightly with salt so the salting out effect must come from G_{u2} (which is negative - the ions don't like to be near the benzene) and, specifically, the Li oddity must come from a less negative G_{u2} than would be expected from such a hard ion. The MD results indicate (but cannot prove because of uncertainties in the force fields) that the strong hydration shell around the Li dilutes interactions with benzene, leading to a less negative G_{u2}.

Everything we have seen about Hofmeister tells us that if we are looking for an ion-solute effect, then instead of looking at ion-water and ion-ion explanations, one should look for ion-solute interactions. It sounds so simple and obvious, but decades of obscure "water structure" thinking show that it has been far from self-evident to the aqueous solubility community.

7.2.5 The KB view of Hofmeister and cellulose

Dissolving cellulose in water should be trivially easy - cellulose is just a bunch of -OH groups which should love water. And, indeed, plenty of polysaccharides are reasonably soluble in water for exactly that reason - though you might have to boil the polysaccharide (starch) to overcome kinetic barriers. It happens that the internal H-bonding network from the specific conformation of the sugar

---

67 Pritam Ganguly, Modeling and Understanding Aqueous Mixtures Using Kirkwood-Buff Theory of Solutions, PhD Thesis, U Darmstadt, 2014

units in cellulose makes it remarkably hard to dissolve both kinetically and thermodynamically. As is pointed out in the key paper on KB and cellulose solubility[68], even good solvents for cellulose need tricks such as "steam explosion" to help the process along, otherwise cellulose can get stuck in "kinetic traps".

Focussing on the thermodynamics, and using cellobiose (the repeat unit of cellulose) as a model system, the solubility in water containing one of a series of salts follows the order $ZnCl_2$ > LiCl > NaCl > KCl. Why is this? There have been the inevitable hypotheses about "water structure" and "hydrophobic effect", none of which makes much sense. Because the solubility of cellulose, u, could be expressed in terms of the standard two terms G_{u2}- G_{u1} as numerator and $1+c_2(G_{22}- G_{21})$ as denominator, it was easy to show that ionic clustering, G_{22} was unimportant as was the ion-water interaction, G_{21}, though interestingly at higher concentrations the whole denominator term became slightly smaller, from 1.1 down to 0.6, suggesting (there was insufficient data to extract the individual G_{22} and G_{21} terms) that the ion-water interactions improve things slightly. What is interesting in this case is that, of course, G_{u2} is highly significant, so that cellulose-ion interactions are important but diminishes at higher salt concentrations while G_{u1} diminishes somewhat less slowly so that the difference becomes more positive. This means that the water is, on average, being pulled away from the cellulose. One might almost say that it is the "water withdrawing power of the salts" at work - taking us back to Hofmeister's original idea. Instead of a vague hypothesis we now have the numbers to show that in this case "water withdrawing" plays a modest but not dominant role.

As the authors point out, both G_{u2} and G_{u1} are negative values, making it look as if neither the salts nor the water much likes the cellobiose. Yet again it has to be pointed out that the large excluded volume effect of that large molecule gets the KBI off to a bad start. In the analysis of solubility we have to be alert both for absolute effects (a large G_{u2} is generally a good thing) and for relative effects (a negative G_{u2} can still be OK in the context of a large solute and an even more negative G_{u1}).

Given the advances (discussed earlier) made in understanding the details of Hofmeister ion interactions with proteins, it seems only a matter of time before we much better understand what is happening with the solubilisation of cellulose and its analogues. KB tells us what needs to be looked at and the tools to do so (NMR, FTIR etc.) are surely available to do the looking.

We also have, as discussed in the COSMO-RS chapter, the clues and insights from the COSMO-RS understanding of how ILs interact with cellulose, with the emphasis on how the hard anions are able to create strong H-bond interactions.

---

68 Thomas W. J. Nicol, Noriyuki Isobe, James H. Clark,and Seishi Shimizu, *Statistical thermodynamics unveils the dissolution mechanism of cellobiose*, Phys. Chem. Chem. Phys., 2017,19, 23106-23112

It is starting to look as if the hitherto intractable problem of what it takes to solubilize cellulose is becoming tractable because we know where to look for answers. For example, because the anion in the KB paper is the constant Cl^-, we have no information on whether the cation effect is (as suggested by the COSMO-RS IL papers) smaller than the anion effect, so a similar set of experiments but with a constant cation would be highly informative. This fits in with a general theme of the book that hypothesis-driven solubility work can be more rewarding than the stamp collecting approach.

7.3 Aqueous polymers

To my surprise, with one exception, I have been unable to find much solubility theory that can be applied to aqueous polymers, other than the issues discussed with respect to proteins and to cellulose. The exception is the confusing world of smart polymers discussed below.

7.3.1 Neutral polymers

Some polymers, e.g. polyethyleneimine or polyvinyl pyrrolidone, just dissolve in water and seem to pose no interesting solubility science problems. As discussed in the KB section, others, such as the starches and celluloses are dominated by internal H-bonding that depends strongly on subtle arrangements of the -OH groups around the sugar rings and are generally insoluble or at least have a strong tendency to gel.

Polyvinyl alcohol (PVOH) is mildly interesting. As it is impossible to polymerise vinyl alcohol (because it exists as its tautomer acetaldehyde) it is made by hydrolysing polyvinyl acetate. The interest arises because partially-hydrolysed PVOH (70-80%) behaves as one imagines PVOH should behave, a boringly water-soluble polymer. But fully-hydrolysed (95%+) PVOH is *very* difficult to work with and can only be put into solution in boiling water. A solution of moderately concentrated high-hydrolysis PVOH gradually gels over time as the strong network of inter/intra H-bonds overwhelms interactions with water.

The aqueous polymer which continues to be studied in detail is polyethylene oxide, PEO (or polyethylene glycol, PEG; the naming convention seems to be arbitrary), which would probably be studied far less if it didn't have the key feature of becoming considerably less soluble in hot water, the reverse of what happens with most other aqueous polymers. The fact that numerous papers pronounce with confidence that *they* have the explanation makes it hard to know what the real explanation is, but it seems to be some combination of loss of the helical structure and the reduction in strength of ether-water H-bonds. The helical structure is said to preferentially expose the ethers to the water and as the structure unwinds more of the hydrophobic methylene groups are exposed.

So we have a bunch of interesting observations about effects in different aqueous polymers, but no general-purpose science that allows us to solve problems. Although I have identified a KB paper on PEO in water, which shows the sorts of trends one might expect throughout the relevant range of water mole fraction, there are no temperature-dependent data from which to disentangle which G_{ij} effects are most crucial to the fall-off in solubility at higher temperatures.

7.3.2 Ionic polymers (Ionomers and polyelectrolytes)

I would have liked to say something interesting about the solubility science of polymers that contain ions. They are loosely divided into two types: the "polyelectrolytes" which are generally thought of as polymers with large numbers of permanent charges, such as sulfonates; the "ionomers" which have relatively small numbers of charges which might themselves be neutralisable, such as carboxylic acids.

However, I have found little that is of general scientific interest, as opposed to fascinating details (important for formulators) with each specific polymer.

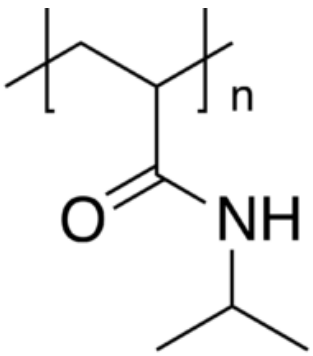
Polyelectrolytes show "typical polyelectrolyte behaviour" which is defined as their relative viscosity increasing sharply at low concentrations. This is not wildly interesting in practice. Simple polymer theory shows that at low concentrations (before there is significant polymer-polymer interaction) viscosity depends linearly on polymer concentration. So the "reduced viscosity" (viscosity/concentration) is a constant. For polyelectrolytes, reduced viscosity increases at low concentrations, meaning that the individual polymer molecules must be stretching out on average and therefore providing a bigger interference with flow of the liquid. The explanation is that at these low concentrations there is no ionic shielding (Debye) of the charge-charge repulsion effects. Not surprisingly the extent of the increase in reduced viscosity depends on how closely the counter-ion (Li^+, Na^+, K^+...) is associated with the sulfonate anion and this usually follows a Hofmeister series.

But because I can think of very few uses for polyelectrolytes at these very low concentrations, I cannot get too interested by this effect. At high (in-use) concentrations the complexities of polymer entanglements and ion-ion interactions are so specific to each polymer that I cannot find any useful scientific guidance.

For the ionomers the key behaviour of interest is how water solubility depends on pH. The widely-used polymers with carboxylic acid side groups are popular because their solubility and viscosity behaviour can be controlled with modest changes of pH. As with the polyelectrolytes, much of the behaviour is fascinating at the detailed level and is vital for real-world use, but I can find no useful scientific principles. We can all say that density of carboxylic groups

and Hofmeister cation effects will be important and we can imagine how these effects play out, but it is hard to say anything beyond those generalities.

7.3.3 Smart polymers



For simplicity I will discuss only one such polymer, the famous PNIPAM, Poly(N-isopropylacrylamide). It is famous not just because its solubility sharply decreases when the temperature is raised, but because the switch happens ~30°C, i.e. not far from physiological temperatures, with claimed implications for all sorts of smart schemes such as drug delivery and switchable nanoparticles. The reported transition temperature changes from paper to paper and depends on subtle details of the polymer and, if it attached to something like a nanoparticle, the curvature and packing density around the particle. The sudden reduction in solubility when the temperature rises above the confusingly named LCST (Lower Critical Solution Temperature) is the opposite to the more normal condition where solubility reduces when temperature falls below the UCST (Upper Critical Solution Temperature). The confusing stupidity of these names (they make sense to thermodynamicists) has been the cause of much pain.

Because PNIPAM has be studied for 40 years, the reason for the change of solubility at ~30°C should be well-known and uncontroversial, yet decades of papers have resulted in "explanations" that have convinced no one not even, I sense, the authors of the many papers. Because the solvent is water, the usual suspects are lined up and confusing tales of water structure, clusters, entropy/enthalpy compensation have been the result. Much of the problem with finding the cause for PNIPAM's behaviour is caused by how the problem has been framed via the language used. If the question is posed in the language of water structure then, automatically, answers that exclude water structure are themselves excluded. In the Language chapter, we see that rephrasing the question about PNIPAM makes most of the puzzles disappear.

7.4 Aqueous solubility summary

If, as the Language chapter argues, we cease to give water any special status and analyse it using the normal language applied to any other solvent, then the solubility issues become more tractable. Neither COSMO-RS nor KB find any need to invoke anything special to explain what is going on, though water's greater ability to dissolve ionic species increases the range of properties that require explanation. I find it significant that exactly the same KB theory copes with the effects of salts and with the effects of neutral molecules on the folding/unfolding of proteins, and that the effects are dominated either by specific interactions or excluded volumes.

The more, therefore, we reduce the mythology of water and the more we apply specific-interaction theories backed up by specific-interaction probes, the faster we will make progress in understanding and mastering aqueous solubility.

8 Green solubility

I believe that the large amount of work in recent decades on green solubility has, on balance, positively damaged the planet. That is a deliberately strong statement and therefore needs some strong justification. Because the use of water as a green solvent is largely uncontroversial (though the 30-50% extra energy needed to evaporate it is not a trivial concern), and because aqueous solubility was discussed in the previous chapter, this chapter is about the choice of solvents other than water.

I am all for saving the planet. My problem is that many greens do not appreciate that resources are limited and that sincerity of motive can often be confused with ability to make a difference. A typical paper reads "Solvent X is non-renewable and toxic so here we describe a bio-based alternative". Before even starting on such a project a simple question should have been asked. "We are going to consume precious resources (time, chemicals, brains...) in this project. If we succeed technically, what are the chances that: (a) 100% replacement of Solvent X would make a significant difference to the planet; (b) we can produce our solvent at an affordable price; (c) we are not trashing the planet in some other way, such as using arable land to grow chemicals; (d) we can afford the few million dollars required to register this new solvent as being safe?"

My view of the countless such projects is that the answers to most of those questions are negative and are predictably negative right from the start[69]. So, by green principles, such a project should never have started. Resources are limited and should not be frittered away on projects that stand no hope of making a significant difference to the planet.

Let us examine each factor in turn.

8.1 Why most green solubility projects will not save the planet

8.1.1 Making a significant difference

If the likely market is, say, 100tpa (ton per annum) then this is so insignificant in terms of the tens of millions of tons of solvents used worldwide that it is not worth even bothering about in terms of saving the planet. The argument that "every little helps" is not appropriate. Every little helps only a little, and what the planet needs is help on the large scale, so precious green brainpower should be focussed on large scale issues. If the likely market is 100,000tpa then we are talking about something interesting but it needs serious industrial muscle right from the start or it will be a hopeless endeavour. Playing around with small-scale

69 I have no problem with cutting edge science that fails because science is hard. My problem is with failure that is predictable from the start. Here the criterion of failure, "not saving the planet", is defined by the request for funding based on the proposition that the project is green.

projects is (with exceptions that need to be thought through carefully) not a good use of precious research resource.

There is another reason why so many green solubility projects will fail - they do not invoke solubility science. Papers say words to the effect of "This bad solvent is hydrophobic so we will replace it with this green solvent which is also hydrophobic". The feeling seems to be that because green is good, work on green is good, even if it defies the laws of physics. A solvent, however green it is, will not make a significant difference if its solubilizing power is inadequate for the job.

As a specific example, the famous solvent FAME (Fatty Acid Methyl Ester, or methyl soyate or biodiesel) is indeed green, it does scale, but it is a remarkably useless solvent for two reasons. First, in HSP terms [16.4, 2.6, 4.5] it is heading strongly in the direction of useless solvents such as dodecane [16, 0, 0], and second because its high MWt (~300) means that kinetically and thermodynamically it is a poor solvent compared to the similar HSP in a smaller solvent such as cyclopentyl methyl ether.

8.1.2 Affordability

In terms of affordability, everyone says that they are "prepared to pay a little more" to be green, but in reality "little" means "very little". I don't know how many readers of this book have ever thought about the scale-up from a nice "green" fermenter in a lab to a 100,000tpa fermentation installation. It is not something for the faint-hearted. Fermentation is a ghastly, energy-intensive, wasteful process in many ways, requiring large amounts of liquid "waste" (full of nutrients etc.) to be processed once the desired chemical has been extracted. Greens love to show pictures of nasty chemical factories, comparing them to lovely fields of whatever is going to be converted into a green chemical. But a large-scale fermenter looks exactly like a nasty chemical factory because that is what it is. Someone has to pay for all this infrastructure and it is a hugely difficult decision if the nasty solvent it is going to replace isn't all that nasty and is already being produced very efficiently from oil.

8.1.3 Trashing the planet in the name of "green"

There is a very interesting report[70] from an EU-funded consortium (ECOBIOFOR) attempting to produce green chemicals for the paint industry. This is exactly the scale opportunity that green solvent efforts should be tackling and they are asking lots of the right question. One of the questions is about how much a green solvent will save the planet. The authors admit that doing a comparison between an entirely new solvent and the current petro-based

---

70 C. Bories et al, *Life Cycle Assessment of bio-based solvents for paint and coating industries*, Setac Poster, on-line https://www.swissbiotech.org/sites/swissbiotech.org/files/textmitbild/dateien/poster_setac16.pdf

version is exceedingly difficult. So, instead, they compared bio- and petro-based butyl acetate. In general the bio-based version was predicted to be modestly better than the petro-based version in many aspects and significantly better in terms of "CO2 storage", i.e. the bio feedstock soaked up CO2. But, of course, the bio-based version was worse in terms of land use and water consumption because the sugar cane was no longer going to feed humans and (I assume) precious water was used on the crops. But now add another factor. How much CO2 will be created by building a butyl acetate fermentation production plant and what are the opportunity costs of spending that money on a new plant (i.e. what more worthwhile green cause could have been addressed with those resources?) when we already have a fully-functioning, highly-efficient conventional production plant? Of course the current plant generated CO2 in construction but that is a sunk cost about which we can do nothing. It is, therefore, not at all clear to me that it is worth devoting resources to this project - though in this case I am assuming that more knowledgeable scientists have concluded that the balance is in their favour. This is one of the rather rare examples of a well-thought-out project with significant resources that, if it succeeds, would make a difference. There are hundreds of research papers on projects that, irrespective of the scientific risk, stand no chance of doing what the papers claim they are doing - saving the planet.

One classic stated source of raw materials for green chemicals is "bio-waste". I have been in a green project that assumed that there was lots of bio-waste sitting around because farmers were unable to get rid of it. As it happens, farmers are not stupid and there is a huge market for dealing with bio-"waste" and for our specific project it was surprisingly difficult to identify a suitable source. The problem is complex. Is it greener, for example, to truck the bulky waste 100s of km to be fermented into a bio-solvent, compared to fermenting it locally into bio-gas to generate power for the local community? And, irrespective of the greenness, which solution is better in the long-term interests of the farmers?

8.1.4 A spare few $ million for safety

The regulatory authorities do not take greenness or naturalness into account when assessing the safety of a new solvent. The solvent has to pass as many safety tests as any other solvent and this means that the up-front costs for a solvent to be sold at scale (and there is little point in wasting resources on solvents that don't scale) are in the multi-million $ range. It just needs one adverse effect on some small organism (or on mice etc.) for the whole of the multi-million $ investment to become valueless. Interestingly, the authorities tend to prefer pure compounds, which make it easier to define and assess the risks, so natural blends (which are typical of many green sources) pose significant extra regulatory difficulties.

The logical outcome of this is that only the largest of corporations, with in-depth internal resources for managing such risks, are going to launch a significant new solvent. Well-intentioned academic or start-up efforts face formidable risks if they want to bring their 100ml demo samples to the 100,000tpa market.

Before switching to a more positive note, let us discuss two other green fads, supercritical CO2 (scCO2) and ionic liquids.

8.2 The hype of scCO2

We have already discussed the science of scCO2 entrainers. The attractions of scCO2 as a solvent (especially for extraction) are two-fold: (1) recovery of the solute is trivial; (2) the low viscosity of the supercritical fluid aids flow through whatever is being extracted. The reason that a few percent of solvents such as ethanol or ethyl acetate is added, complicating the simplicity of the first attraction, is that the low density and the molecular structure of CO2 makes scCO2 a rather poor solvent - on a par with something like isopentane.

I found it very depressing to have to read countless scCO2 papers which basically say "scCO2 is green, and therefore good, so we will try to use it on this totally random solute". It was equally depressing to read (not in exactly these words) "we have no good ideas of how to model solubility in scCO2 so we will use this random approach that seems to work". The net result is that there is plenty of data but little understanding. Once again the greenness has been used as an excuse to do science that takes us nowhere. This seems to me to be a squandering of precious research resource.

My summary of the conclusions from those decades of scCO2 research is: you will only get significant solubility if the ideal solubility is high (i.e. low MPt solutes) and if there is the chance of significant H-bonding (donor/acceptor) interactions between entrainer and solute. The one entrainer that could have solved many scCO2 problems, water, suffers from the drawback of being so insoluble in scCO2 that you cannot (in general) add enough of it to make a big difference.

scCO2 with a suitable entrainer can be an excellent green choice. The upsides need to be weighed against the downsides of the complex systems for handling the high pressures and temperatures. In cases like caffeine extraction scCO2 does a good job. The point here is not that scCO2 is bad (it isn't) nor that research on scCO2 is bad (it isn't). The point is that doing lots of research on scCO2 "because it is green" is a profoundly un-green activity because resources are precious. The probability of scCO2 being the best option is low, so precious resources should be used only when there are compelling reasons to suggest that it is a method superior to the alternatives.

8.3 The hype of ionic liquids and NADES

Some years ago I became increasingly irritated with the non-stop praise for the amazing ionic liquid (IL) green solvents based on, say, imidazoles and PF_6^- anions. These seemed to me to be hopelessly expensive and probably highly toxic. And the MWt of just the anion (145) is higher than the MWt of most solvents so the molar and thermodynamic efficiency is off to a bad start. But what did I know? By chance I met a younger expert in ionic liquids and asked about this peculiar situation. He told me that he and many others had been equally bemused, but when Professor Famous gave yet another talk about these wondrous green solvents, no one had the courage to openly argue against them.

Fortunately those dark days are past and such claims are seldom heard. Yet the hype over ILs continues. I would not have used the word "hype" in this book without the existence of a splendid review[71] from Kunz with a title using exactly that word: "The hype with ionic liquids as solvents".

It is undoubtedly the case that ILs have some remarkable properties which make them suitable in niche areas where alternatives are lacking. Anyone who has tried to dissolve cellulose can only marvel at how good some IL systems can be, and I appreciate that by abandoning absurdly expensive cations and nasty PF_6^- anions replacing them with, perhaps, rather simple, affordable acid/amine ILs there are serious commercial opportunities for them. Whether or not they are green is largely irrelevant; the key is that they can sometimes do a job that other solvent systems cannot do. If work on ILs had started by recognising the multiple disadvantages of ILs and therefore focussed on tasks for which they are spectacularly good I would have no problem with them.

Beyond a small number of special uses, the case for ILs being green seemed to rest largely on the fact that they are non-volatile and, as we all know VOCs (Volatile Organic Compounds) are bad for the planet. But "bad" is only relative. We have an excellent infrastructure for handling VOCs without endangering the planet; for example the fact that they are volatile makes them easy to recycle via distillation, and as a last resort, if there is a large quantity of mixed-up (chlorine-free) VOCs then they can be used as fuels in, for example, cement factories. Given that the properties of ILs are often strongly-dependent on low levels of moisture, the classic technique of precipitating ILs with water then drying them thoroughly is not without its environmental impact. And in any case, the high viscosity of these ILs make them useless for many processes.

---

71 Werner Kunz, Katharina Häckl, *The hype with ionic liquids as solvents*, Chemical Physics Letters 661 (2016) 6–12

But what makes the whole IL saga seriously non-green is exemplified by a review[72] on the biodegradability of ILs. First, and not surprisingly, many of these "green" solvents are poorly biodegradable. It is well known that the popular imidazole group seems to be alien to most bacteria and remains untouched. As the review points out, biodegradability is only a first step - what about ecotoxicity? More significantly, the review covers ~300 ILs. This means that groups of researchers around the world have synthesised at least 300 ILs not with some specific aim of finding an IL to achieve a specific property; instead they have been synthesised because ILs are green and green is good. No one seems to have asked by how much the planet has been trashed by all the chemicals and energy used to make and test these molecules, most of which stand no chance of solving the problems which is the excuse for creating them in the first place. Nor has the scientific community questioned the opportunity cost of all this work. What would have happened if these well-intentioned scientists had worked on issues that stood a significant chance of saving the planet?

The Kunz hype review provides these wise words: " All in all, ILs, proposed as solvents, seem to be of much less importance than the huge number of papers might make us believe. In our opinion, they should be considered as high performance chemicals for special applications rather than as promising alternative solvents for common chemical and industrial processes."

8.3.1 Natural Deep Eutectic Solvents (NADES)

The hype around ILs has somewhat abated, in time for a new round of hype on NADES. The name says that you mix two natural solids (e.g. choline and glucose or choline and lactic acid) and, amazingly, a liquid (sometimes ionic, sometimes not) is produced at the eutectic point which is very much lower (i.e. deep) than the melting points of the solids. NADES are a subset of the more generalised deep eutectic solvents (DES).

Again, I have no doubt that these wonderful systems will have some outstanding properties which will allow them to create novel solutions to difficult problems. I am all for that - provided that scientists pursue the science in that direction. What is not acceptable is the blind hype that these are green and natural and therefore will be our next green solvents. They will not be. They are way too expensive for most uses, take away natural resources that could be used for other purposes, they are usually far too viscous to be useful, and there is plenty of data to show that many of them show significant toxicity.

This last point is a surprise to many. NADES papers tend to start by saying that each component is safe so the mixture will be safe. That is not how toxicity works. And when you start to think about why something like choline exists, it is

---

72 Andrew Jordan and Nicholas Gathergood, *Biodegradation of ionic liquids – a critical review*, Chem. Soc. Rev., 2015,44, 8200-8237

precisely because it has very powerful solubility (or, better, solubilizer) behaviour within natural systems. And anything which is "powerful" in a natural system can be both good and bad. So it is no surprise that a review of the cytotoxicity of NADES[73] points out that many of them show significant toxicity. A more recent review[74] includes the following in the Abstract: " NADESs exerted cytotoxicity by increasing membrane porosity and redox stress. In vivo, they were more destructive than the DES and induced liver failure. The potential of these mixtures was evidenced by their anticancer activity and intracellular processing." That is not the sort of quote you would want to see attached to your green press release. And remember that "anticancer" often means "is really good at messing with DNA", so one had better be sure that the DNA that is being messed with is only the bad DNA in cancer cells.

The high viscosity of NADES encourages the use of diluents, which in practice means water. Unfortunately it seems to be the case that by the time you have added enough water (say 10%) to lower the viscosity, the solubility properties of the mixture are no longer so exciting. Presumably whatever creates the strong intermolecular links that produce the high viscosity is responsible also for the solubility power for the solutes of interest.

I have not been able to find any reports about the skin irritation potential of NADES. But the modest solubility of wool keratin in choline/urea suggests to me that these are not solvents I would generally like to have anywhere near the keratin that makes up the surface of my skin. It is truly impressive that ILs and DES can dissolve keratin; at the same time it is a reminder that these are likely to be vicious solvents in the hands (literally) of the careless.

8.4 Being positive about green solvents.

In my view the only companies that will make a significant impact on green solubility issues are the large chemical companies. Given that the default view is that "large chemical companies are automatically bad", this might sound surprising. Yet a few moments' thought shows that it is the only way that significant changes *can* be made.

First, only those solvents used on the large scale are going to make a significant difference to the planet, and only large companies have the capability of making the investments required not just to produce a lot of chemical but to do it efficiently with minimum waste. However cynical one might be about chemical companies, their accountants do not like lots of wasted energy or chemicals.

---

73 Maan Hayyan, Yves Paul Mbous, Chung Yeng Looi, Won Fen Wong, Adeeb Hayyan, Zulhaziman Salleh and Ozair Mohd-Ali, N*atural deep eutectic solvents: cytotoxic profile*, SpringerPlus 2016, 5, 913/1-913/12

74 Yves Paul Mbous, Maan Hayyan,, Won Fen Wong,Chung Yeng Looi,and Mohd Ali Hashim. *Unraveling the cytotoxicity and metabolic pathways of binary natural deep eutectic solvent systems*, Sci Rep. 2017; 7: 41257

Second, a new green solvent needs the investment of $millions to get through regulations such as Reach. No green start-up can ever risk those sorts of sums when it needs just a few dead bugs, fish or animals to make that whole investment worthless.

Third, end users need a lot of convincing to change from their known less-green chemical to an unknown green one. This requires lots of sales and marketing resources plus the ability to sustain the inevitable early losses because the selling price needs to approach that of the established alternative, making it impossible to quickly cover the up-front costs of developing the new solvent.

Well-intentioned efforts from universities and small companies to come up with a blockbuster new green solvent are, therefore, likely to fail. And if these smaller institutions focus on low-volume niche solvents, the impact on the planet is so small as to question whether it is green to devote precious resources to such a target. Small-scale efforts to produce niche solvents that are profitable green-wash are to be analysed like any other business proposition. The chief issue will be whether the quantities are sufficiently small to not require the full-scale tox testing that soaks up the $millions.

So the sad, realistic news is that as formulators we have little chance of becoming greener by inventing or gaining access to amazing new green solvents. So we have to go greener by being smarter with what we have.

8.4.1 Being smart about green solvent options

You can imagine that I get easily irritated by papers on green solvents. I was, therefore, delighted by a major (open access) review paper[75] about green solvents for pharma. This is an industry that needs to look and act green and there has been extensive cross-industry work in order to establish best practice for solvents. Because these are hard-headed industrialists their concern is with "what will actually help the planet", not "what fashion says will help the planet". So their first two priorities were to identify those conventional solvents whose manufacture had the minimum impact on the planet and where the overall impact on human health and that of the wider environment was minimized. However, these priorities had to be tempered by the fact that, like it or not, many pharma compounds are insoluble in the solvents we tend to think of as nicer. So they also had to select the least bad solvents from really useful pharma solvents such as DMF, NMP or DMSO. Why this pragmatism rather than devote resources to finding all-green perfect solutions to their solubility problems? Because the impact for good on the planet by this large industry switching *now* to a set of better solvents is huge. The probability of provably better solvents

75 Fergal P. Byrne, Saimeng Jin, Giulia Paggiola, Tabitha H. M. Petchey, James H. Clark, Thomas J. Farmer, Andrew J. Hunt, C. Robert McElroy and James Sherwood, *Tools and techniques for solvent selection: green solvent selection guides*, Sustain. Chem. Process., 2016, 4, 1–24

coming along any time soon is low and the resources required to find them are arguably better spent on other priorities.

The "environmental impact" of the manufacture of conventional solvents includes whether a given solvent is created efficiently within an overall manufacturing ecosystem. A stand-alone plant to produce any single solvent (green or otherwise) may not make much environmental sense compared to production of, say, acetone, within a highly-optimised site that also uses acetone as a solvent or raw material in other processes.

A more general-purpose recommended list of conventional solvents can be found in the open access Chem21 report[76]. The usual alcohols are fine, most of the usual ketones and esters are fine, the best ether is anisole, for hydrocarbons heptane and toluene are reluctantly accepted as there are no alternatives and of the aprotic polar solvents, the ones crucial for pharma, acetonitrile and DMSO are the least bad. The pharma industry is split about whether sulpholane is good or bad - it is one of those interesting judgement calls.

The same 2016 review also looks at some promising "unconventional" (their term) solvents. I am surprised at how many solvents generally assumed to be green, such as ethyl lactate, γ-valerolactone and glycerol (both of which are food safe) are flagged as "problematic" in the review. One reason is that recovery of high BPt solvents is environmentally troublesome if the default method of distillation is used. An alternative to toluene, p-cymene is suggested, though its overall score is not that much better than toluene's. Armed with the thorough comparison of potentially greener solvents, the user can make an intelligent choice, including factors such as price, to arrive at an overall best-practice decision.

Is this the perfect way to save the planet? No. Is it the smart way? Yes.

8.4.2 Being smart about green solvent choices

The logic of the previous section allows you to know what solvents you might possibly use. Now you have to work out which ones to use and that, of course, depends on the solute.

Again it is frustrating that so much of the green movement wastes precious resources by not formulating using rational solubility tools. In a large organisation concerned with pure solutes and with access to COSMO-RS, finding a rational solvent or solvent blend is straightforward. The predictive power is so great that it takes relatively little further optimisation to achieve the

---

76 Denis Prat, Andy Wells, John Hayler,Helen Sneddon,C. Robert McElroy, Sarah Abou-Shehada and Peter J. Dunn, *CHEM21 selection guide of classical- and less classical-solvents*, Green Chem., 2016, 18, 288–296

target - assuming that the ideal solubility of the solute is large enough for the approach to work.

For general formulators working with more complex materials it makes sense to measure the HSP of the target (using small amounts of not-necessarily-desirable solvents) and then to use a tool like HSPiP to find a suitable optimum solvent or blend. As a specific example, here is the U York Green Chemistry Centre of Excellence's list of green solvents taken from HSPiP. The list includes the MVol (because smaller is generally better) and the RER, Relative Evaporation Rate, because that can have a big effect on the choice of solvent.

Solvent	δD	δP	δH	MVol	RER
Water	15.5	16	42.3	18	80
Ethanol	15.8	8.8	19.4	59	150
2-Propanol	15.8	6.1	16.4	79	150
1-Butanol	16	5.7	15.8	92	43
t-Butyl Alcohol	15.2	5.1	14.7	96	160
Isoamyl Alcohol	15.8	5.2	13.3	109	17
Acetone	15.5	10.4	7	74	560
Methyl Ethyl Ketone (MEK)	16	9	5.1	90	380
Ethyl Acetate	15.8	5.3	7.2	99	390
Isopropyl Acetate	14.9	4.5	8.2	118	350
n-Butyl Acetate	15.8	3.7	6.3	133	100
Anisole	17.8	4.4	6.9	109	17
Glycerol	17.4	11.3	27.2	73	0
Dimethyl Carbonate	15.5	8.6	9.7	85	195
d-Limonene	17.2	1.8	4.3	163	12
Ethylene Glycol Diacetate	16.2	4.7	9.8	133	1.4
p-Cymene	17.4	2.27	2.44	156	13
Cyrene	18.9	12.4	7.1	102	2.7

This list of 18 solvents is an explicitly "purist" list, excluding many of the pragmatic compromises from the Chem21 list. So it is unlikely that any single solvent will be a great match for a given solute. Taking PLA as a "green polymer" example, HSPiP shows us that anisole is the closest match, with a distance of nearly 4. How can we do better? An automatic search shows that a 55:45 blend of anisole:cyrene is a near-perfect match. This blend may or may not appeal to you as there are many other factors such as cost and volatility that must be taken into account. What is important is that via a rather straightforward process involving a list of solvents that have been objectively judged as being green, the user can generate rational formulation options that can lead to an efficient (and therefore green) use of lab time to obtain a real-world formulation.

HSPiP contains a much larger green dataset selected by the Aubry group in Lille using a different set of explicit criteria. With this much larger list it is easier to find a single solvent or blend to achieve your specific goal, but the list is more idealistic (it is a scientific list so that is not a criticism) so the availability and affordability of the optimised solvent(s) has to be examined carefully.

8.5 The take-home message about green solubility.

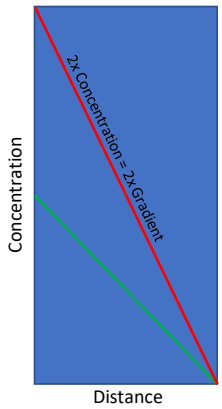
There are many fine scientists doing excellent work on saving the planet via green solubility. Unfortunately there is also a deluge of well-intentioned work under the green banner which has been a waste of precious resource, the opposite of what their intentions were supposed to be. It is easy to criticise things in hindsight. But the deluge of papers on ILs could easily have been avoided by the application of a little common sense. A greater focus on solubility science would have alerted the community much earlier on about the solvent properties which might be required by end users in order to substitute for the worst-offending, large tonnage solvents. An earlier focus on the likely trade-offs of using a bio-resource for generating a solvent compared to that resource being used for food or, say, bio-gas would have quickly eliminated many projects. And much effort could have been saved if researchers appreciated the near-impossibility of going from lab idea to million ton green solvent sales without access to the risk capital needed to prove safety and efficacy of the new solvent whilst producing it at an affordable cost. Finally, for those intending to produce a few tons of a niche solvent, the impact on the planet is so irrelevant that precious resources should not have been consumed in the name of greenness.

Saving the planet is hard. It needs focus, huge concentrated resources and good science. My view of much of the work on green solvency is that it has lacked focus, has squandered large amounts of modest resources and has all-too-frequently ignored basic solubility science. We could have, and should have, done much better than that because resources are both limited and precious.

9 Diffusion and Solubility

If you want to avoid a chemical going through your protective glove, if you want your flavour molecules to stay in your food and not escape through the packaging, if you want the solvent in your coating to have the right level of "bite" into your substrate surface then you need to understand diffusion science and the key part that solubility science plays.

There are just three solubility-related facts you need to know about diffusion. For simplicity I will use the words "solvent" and "polymer" to describe the thing that is diffusing and what it is diffusing through. The same principle works whatever diffusing species is going through whatever material, until the material is so open and porous that the discussion shifts to permeability (Darcy's law) which is not covered here):

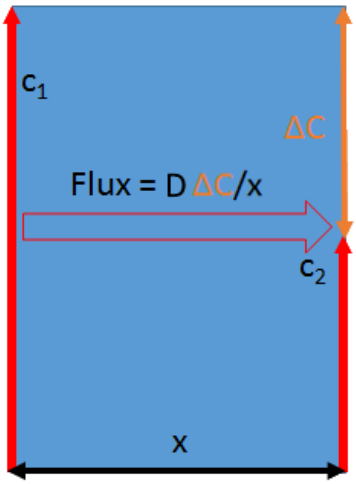
1. 

The first is that the rate at which the solvent diffuses across a given thickness of, polymer is proportional to the concentration gradient across that thickness. Suppose we have a diffusion situation where everything that gets through the polymer immediately disappears so that the concentration of solvent in the top few nm of polymer at the right side of the polymer is zero. Now compare the situation where the concentration in the top few nm of polymer on the left size is doubled (red) compared to the other case (green). The rate at which the solvent will diffuse through that thickness of polymer will be doubled. If you decided to double the thickness of the polymer in the hope of providing a better barrier, then the rate of diffusion would halve for both cases. So if you want to control diffusion you must control the concentration in the top few nm of the relevant surface. Why "the top few nm"? We have to define a concentration on one side in order to know the concentration gradient and thinking about it as the amount of solvent in a physically significant amount of polymer makes sense. If you try to think of it as "solubility" in the top few molecules it is harder to relate to standard solubility thinking.
2. The second key fact is that for a given concentration gradient, the rate at which the solvent moves through the polymer is proportional to the diffusion coefficient. Surprisingly (at first) this has no relation to solubility. It depends, of course, on the polymer. It is generally accepted that nothing much diffuses through crystalline polymers, and polymers with freely-moving chains will have higher diffusion coefficients than those with rigid chains or with highly crosslinked structures. The diffusion coefficient also depends strongly on the shape and size of the diffusing species. Small spherical or flexible linear molecules will diffuse faster than large, complex or rigid molecules. Diffusion is a random process where molecules will hop to the next available sufficiently large, empty space. It becomes steadily harder to find an empty space as

the molecule gets larger. As we will see, the diffusion coefficient of a 10nm nanoparticle in PE is 23 orders of magnitude lower than a typical solvent molecule with a MWt~100.

3. Having acknowledged that solubility does not affect the diffusion coefficient (so an "insoluble" molecule will diffuse at exactly the same rate as a "soluble" one), the third key fact is that solubility has dramatic effects on diffusion coefficients. At the crucial outer surface in contact with the solvent, if the solubility is very high then the polymer will be highly swollen with solvent. This means that it is much more open and flexible, so the diffusion coefficient will be much higher. As the solvent diffuses deeper into the polymer the molecules at the solvent front will be moving at the slowest rate because they are diffusing through pure polymer. As more solvent arrives, the polymer swells, so the diffusion coefficient increases. Whereas low-concentration diffusion can be modelled with simple algebra, concentration-dependent diffusion requires numerical modelling, especially when the effects of concentration on diffusion coefficient are large. For typical polymers with good solvents, the diffusion coefficient can increase by three orders of magnitude.

9.1 Fickian diffusion



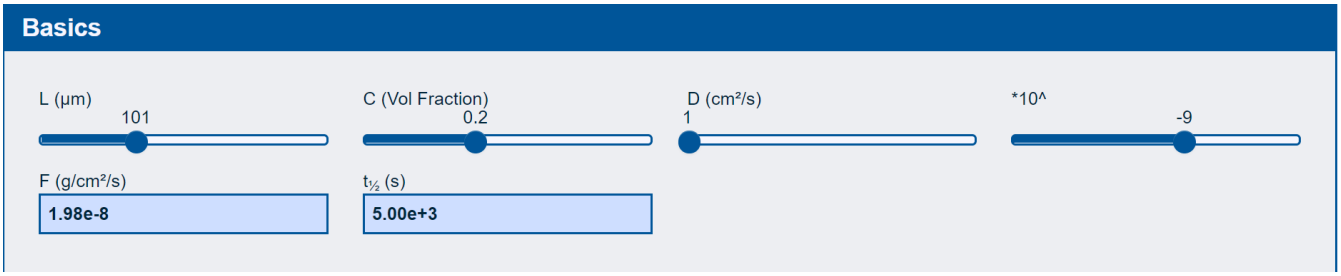
The graphic tells us the basic Fickian diffusion law, that the flux (e.g. in g/cm^2/s) equals the diffusion coefficient, D, times the change in concentration, ΔC, divided by the distance, x, over which the concentration change is measured. We are often interested in the change of concentration with time, ΔC/t and that is just the change of flux over distance, or Δ(DΔC/x)/x. Put these into standard calculus notation and:

$$Flux=D.\delta C/\delta x$$

$$\delta C/\delta t=D.\delta^2 C/\delta x^2.$$

Because these are differential equations which most of us find hard to solve, I've implemented them as apps.

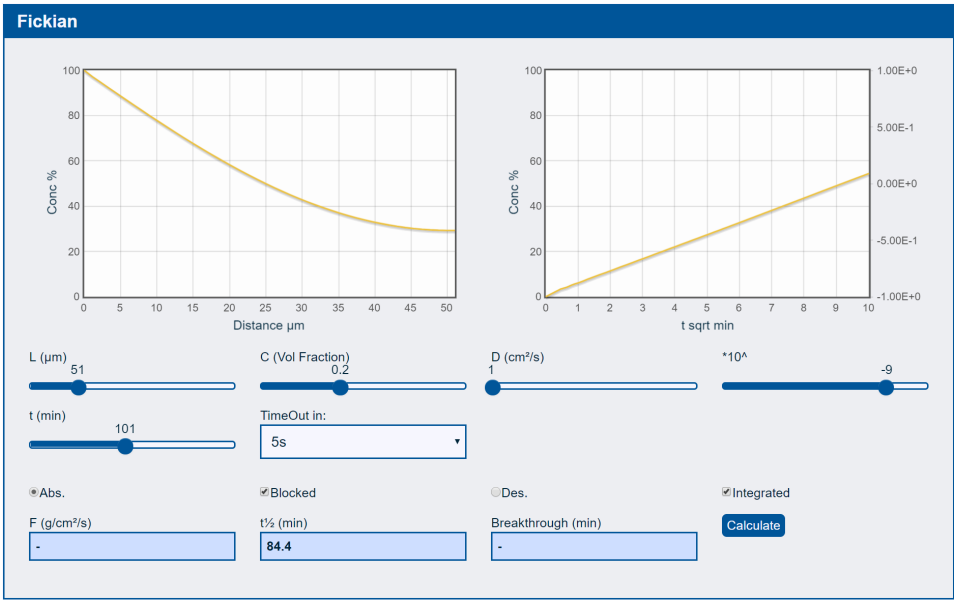
The first app calculates the flux for a given concentration gradient and diffusion coefficient and uses the one common algebraic formula that is useful which tells us the half-time to fill an empty polymer with solvent or for a filled polymer to empty which is $t_{½}=0.049x^2/D$.



App 9-1 https://www.stevenabbott.co.uk/practical-solubility/diff-basics.php

Throughout my apps I use the units cm^2/s for D. I normally choose proper SI units, but somehow the majority of the diffusion world work in these units and $g/cm^2/s$ is a comfortable unit for the amount permeating. A typical D value for a liquid is 10^{-5}, for a small molecule in a polymer is 10^{-9} and for a nanoparticle in a polymer it is10^{-30} cm^2/s. The half-time to fill or to empty (from both sides) a 100µm sample of polymer of the small molecule is 5000 s (~3 days) and for the nanoparticles that is 5.10^{24}s which is far longer than the age of the universe.

To see Fickian diffusion in action, the full app has lots of options. You can model adsorption and desorption and in adsorption mode you can block one side so that it fills up or have the other side at zero concentration to model permeation. You get the flux at the endpoint of the simulation and if you select the Integrated option you can see how much has flowed in or out.



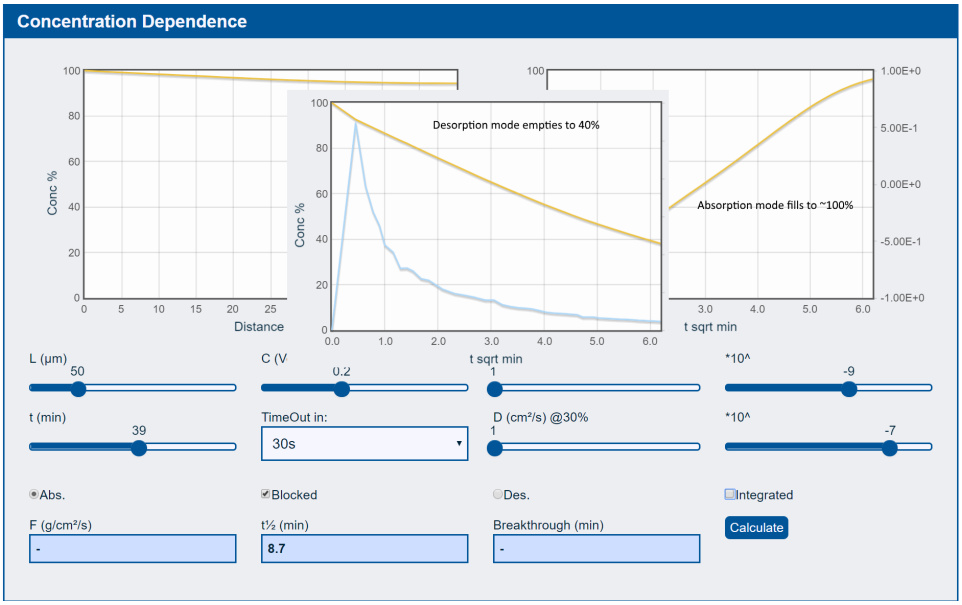
App 9-2 https://www.stevenabbott.co.uk/practical-solubility/diff-fickian.php

The two graphs show different aspects. The one on the left shows the concentration through the sample relative to that at the outside (left side) defined as 100%. This is a 50µm sample and the half-time to fill it is 85min=5000s from the previous app. The previous app had a 100µm sample but was being filled from both sides. The graph on the right is the Integrated view of the amount that has entered. Note that the x-axis is the square root of time because Fickian

diffusion is linear in t½. There are no values in Flux and Breakthrough because those only apply to the unblocked system.

I hope that the app shows that from a couple of very simple formulae a lot can be calculated that would not be evident by just looking at the formulae.

The next app reveals the impact of concentration dependent diffusion. It is not so much the obvious point that you get more rapid diffusion but a little-appreciated consequence of the effect.



App 9-3 https://www.stevenabbott.co.uk/practical-solubility/diff-cdepend.php

Here I have superimposed the desorption right-hand graph on top of the absorption simulation. In absorption mode the same setup as before is completely full after 39min because the diffusion coefficient is concentration-dependent, increasing from 1.10^{-9} at low concentrations to 1.10^{-7} if the concentration ever reaches 30% (in fact it saturates at 20% in this example). Without the concentration-dependence, it would have filled (you can test this in the app) to only 34%.

If it fills up in 39min surely it should empty in 39min. But clearly it doesn't - it has emptied only to 40%. Why is this? Imagine that the left-hand edge of the sample is a tap (faucet). Because in absorption mode the outer few nm are saturated at 20%, the diffusion coefficient is large so the tap is open fairly wide and the solvent can rush in to the sample. In desorption mode, the outer few nm are in equilibrium with the outside so the diffusion coefficient is low and the tap is almost closed. It is therefore difficult for the solvent to escape.

This phenomenon once angered an important Asian customer for a UK-based business in which I worked. It was only when Charles Hansen (a world expert in diffusion science) taught me about concentration dependent diffusion that

I understood what had gone wrong. We had produced an excellent coating onto a polymer. We deliberately used a solvent that gave enough "bite" into the polymer to give us good adhesion. We used HSP to get the blend just right; too little bite and the adhesion was poor, too much bite and the polymer surface became damaged. We had excellent drying capabilities and when the coated product exited the machine it was 100% solvent free. The key solvent was cyclohexanone which has a pungent odour which you cannot miss. The product was shipped to Asia (so it took a while to get there) and when the customer opened it there was a strong odour of cyclohexanone which made them justifiably angry.

What had happened was that the cylohexanone had gone very quickly into the polymer and we had assumed that it had come out equally quickly. But because of the concentration dependence, some of the solvent got trapped inside. During the time it took to travel to Asia it slowly escaped from the polymer and got trapped inside the packaging material, ready to be released on opening.

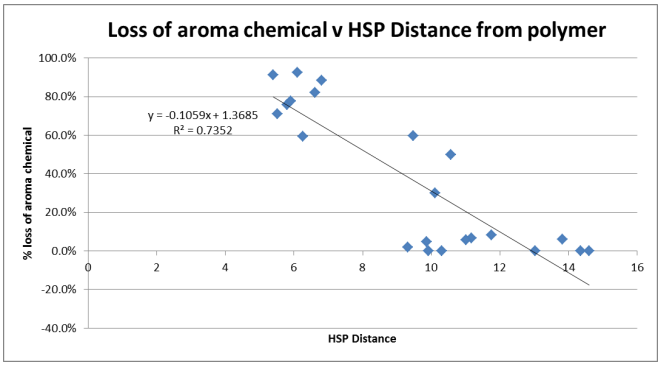
The (solubility) reason for how it got trapped inside the packaging will be discussed in the next section.

The app makes clear a phenomenon that would otherwise seem very obscure. Those who know the phenomenon of "skinning" in coatings will also be familiar with a consequence of concentration-dependent diffusion. As the surface layer of a coating starts to dry out, the rate of diffusion of the solvent through the surface decreases and the solvent in the lower, still fluid, layer gets trapped. If the drying oven temperature exceeds the boiling point of the solvent then the rapid expansion of liquid to vapour can cause a blister in the coating.

9.2 Practical implications

9.2.1 Flavour scalping

Even the simplest "all natural" fruit flavouring is a mix of 20 chemicals. The subtle balance of these flavour molecules is the art of the flavour and fragrance specialist. If somewhere down the line that delicate balance is upset by some of the, say, less polar molecules escaping through the packaging then the flavour of the final product will be unacceptable. The loss of flavour molecules is called "flavour scalping".



When I got a phone call to help solve a flavour-change problem I had no idea whether I could help or not as it was way outside my expertise. I was sent the list of flavour molecules with the % of each of them that got lost over a few days within a package,

plus I was told what polymer was being used. Because diffusion theory says that the rate of loss depends on the equilibrium concentration in the flavour side of the polymer, and because I wanted to try the simplest idea first, I plotted the % loss versus the HSP Distance[77], expecting to see some correlation showing that a small distance implied a larger loss. When, with no further refinements, I got the correlation shown in the diagram it was unnecessary to do any further work, the root cause was clear. The only ways to improve the situation were to add some barrier coating (impossible under the circumstances) or to change the polymer (also impossible). So the project terminated. Had the team known of simple diffusion theory before they started, they would quickly have identified this issue and either stopped the project straight away or gone down a different packaging pathway from the start.

To show how unfamiliar this basic science was to the team, they asked me where the flavour molecules had gone, given that there was no aroma on the outside of the package. I said that the molecules would be sitting in the polymer of the package. They laughed, because they knew (somehow) that this was impossible. However, when they put samples of the package into a headspace GC they quickly found all their lost molecules. In defence of the team, before they had come to me they had consulted some major food experts who had been equally baffled by this issue and who were equally unaware of basic diffusion science.

Given the limitations of HSP theory it is not surprising that the correlation in the graph was adequate rather than excellent. As mentioned in the COSMO-RS chapter, it is possible to estimate flavour-in-polymer solubilities and partition coefficients[78], so for those who are seriously into flavour barriers, COSMO-RS might be a better tool to use.

While on the subject of food, it is interesting to note that at the time of writing, there are just two common ways to create transparent high oxygen barrier food wrappings (aluminium barriers are superior but not transparent). The fancy, complex way, is to use AlOx or SiOx, super-thin vacuum deposited layers of aluminium or silicon oxides. Thick layers are really good barriers until they are handled, in which case they crack and become poor barriers. Thinner layers are resistant to cracking but can have pinholes, making them less perfect as barriers. Because the oxides are so delicate, a protective coating or extruded layer has to be added. The simpler way to make an oxygen barrier is to co-extrude PE-EVOH-PE with the thinnest possible layers (typically 25-4-25μm) compatible with the practical needs of the food packaging. The EVOH multilayer

---

77 For each chemical I could find the SMILES formula and HSPiP automatically estimated the HSP of each flavour molecule and knowing the HSP of the polymer (from HSPiP's database) the distance could be calculated. Thanks to automatic "name to SMILES" websites, the whole process from receiving the flavour names to having the graph shown here took less than 1 hour - for a problem that flavour experts had puzzled over for weeks.

78 Here is the reference again: Christoph Loschen and Andreas Klamt, *Prediction of Solubilities and Partition Coefficients in Polymers Using COSMO-RS*, Ind. Eng. Chem. Res. 2014, 53, 11478−11487

system is in most cases just as good as the fancy system. The reason is that the HSP distance from O_2 to EVOH is large, so its solubility in EVOH is small, so its rate of diffusion is low. For reasons discussed below, the PE-EVOH-PE combination is an excellent general-purpose barrier film. What doesn't get stopped by one polymer gets stopped by the other one.

If you happen to want a polymer film that lets O_2 through at a high rate ("breathable packaging"), choose a silicone. The HSP distance to oxygen is very small so the concentration gradient driving the diffusion is relatively high.

One area where the food industry excels in terms of diffusion science is testing for whether anything nasty can come out of the packaging material into one of a small number of food simulants (oil, dilute ethanol, acetic acid...). Sophisticated computer codes exist for such calculations and we will see how we can all benefit from their work thanks to the food world's formulae for estimating diffusion coefficients in standard polymers.

Finally, the reason that the cyclohexanone got trapped in the packages shipped to Asia is that the HSP Distance of cyclohexanone to PE (the simple packaging film) is large so diffusion out of the packaging was very low.

9.2.2 Safety clothing

If you had to advise a fire-fighter tackling a major spill of a chemical what protective suit to wear, or if you want to choose the right pair of gloves for handling a toxic chemical, or if you want to design a suit that is light, comfortable and resistant to nerve gas, or if you wanted a membrane beneath your house to stop natural radon gas (from the granite rocks beneath your home) from leaking in, how would you go about choosing the right material?

The specific answer to each of those questions is to calculate the HSP distance from the chemical to the relevant materials and choose the ones with the largest distance.

It turns out that in nearly all cases there is a trade-off between finding a general-purpose material and having large HSP distances for all relevant challenge chemicals. If you look at the HSP distance of radon from typical polymers then the obvious choice for a barrier is polyvinylalcohol (PVOH) or polyethylenevinylalcohol (EVOH). As these polymers are easily destroyed by water they cannot be used between soil and a house. Polyethylene (PE) is totally resistant to water but is a poor radon barrier as the HSP distance is much smaller. If you make a membrane with, say, 250μm of PE, 5μm of EVOH and another 250μm of PE you have a tough, practical, excellent barrier to radon and only barriers of this sort of design are approved for radon resistance.
If you play with the standard Fickian app you find that thickness plays a far smaller role than the concentration in the outer nm. Via HSP we know that

these concentrations can differ by 3 orders of magnitude. That is why just a few μm of EVOH is all that is needed to be a good barrier. The layer could be even thinner, but then the chances of a pinhole-free layer become smaller. It is not a coincidence that the basic construction of a radon barrier is the same as standard oxygen-barrier food packaging which is also a general purpose flavour barrier.

Fire-fighters can now enter the name or CAS number into a smartphone app (for which I wrote the calculation "engine" while app experts wrote the interface) and the app will select the appropriate garment from their (limited) choice, based on an HSP-based calculation of how long it would be safe to wear each garment for the time (hours) the firefighters might be exposed. It is very much *not* a secret that the app is based on code more-or-less-identical (though more complex because of the multiple layers of materials) to the Fickian diffusion app. Where safety is concerned, the relevant authorities need to understand the chain of logic behind it.

Exactly the same process has been used to ensure that the right fabric can be used for chemical warfare agents. Because multiple thin layers can be much more effective than a few thick ones, modern suits can be light as well as safe. It is common knowledge that something like PE-EVOH-PE is a good starting point for designing such a suit, though given the super-high risks involved, real suits require more sophisticated structures.

The ProtecPo program (https://protecpo.inrs.fr/ProtecPo/jsp/Accueil.jsp) allows users to find the right gloves for resistance to any of the 1000s of chemicals in their database or blend of those chemicals. The project carefully re-measured the HSP of all the key types of gloves on the market (Hansen had determined many such values some decades ago) and calibrated the Distance-versus-permeation values for good safety. When the user chooses a solvent (or solvent blend), the HSP is calculated and the gloves with the highest Distance are recommended, or, if there is no glove with a very large Distance, a compromise, short-exposure glove is suggested.

9.2.3 Adhesion

Contrary to common belief, practical strong adhesion does not depend on surface energy. Instead it requires intermingling or entanglement across the interface so that crack energy becomes dissipated by polymers slipping and stretching. So during any coating process, it is vital that the solvent can diffuse sufficiently into the coated surface in order to open it up enough for the other polymer (which must, of course, be nicely soluble or swollen) to diffuse among the polymer that makes up the coated surface.

This more or less repeats what was said in the HSP chapter, and the Helfand formula is also necessary to complete the story. The diffusion science interest

comes from a phenomenon that has caused many problems to many printers in many industries. It is often necessary to print/coat multiple layers. Here we will take the most common example which is colour printing. Very often there will be an all-over white ink, followed by the various CMYK and specialty colours. The curious phenomenon is that each ink on its own might adhere well, tests on the printed stack of colours will all be passed easily until one more layer is printed and the adhesion fails catastrophically.

Like many others I used to wonder what was special about the specific layer that caused the problem. And very often, if the print order is changed, that layer causes no problems at all. It is very confusing.

Now recall my embarrassing problem with my Asian customer. Suddenly a new cause becomes a possibility. Each ink layer is printed with a solvent that is designed to have not too much and not too little bite into the layer below, allowing the intermingling and entanglement required for good adhesion. Each layer is carefully dried because residual solvent would be a problem. However, at each step, some solvent will go in quickly and come out slowly. So with each extra layer there might be a little more solvent migrating within the stack. The solvent always diffuses to a zone with the lowest concentration, i.e. towards the interface with the substrate. If the solvent is not especially happy in the substrate, then its concentration will build up at the interface. And if, after enough printed layers, the concentration builds up high enough, that interface becomes weakened (swollen) by the residual solvent and adhesion fails.

This combination of solubility (HSP) and diffusion science is a potent way of troubleshooting adhesion problems.

9.2.4 Controlled release

If an API (active pharmaceutical ingredient) is placed in a safe, bio-compatible polymer such as PLA or PGLA (poly (lactic-co-glycolic acid)) then two potential problems are solved. First, if the API's solubility is too low for delivery in conventional form this provides a way to place where the API can become bio-available. Second, instead of a sudden, large dose of soluble API, there can be a steady ("controlled") release of API over time for long-term action.

The solubility profile of the API in the polymer affects the controlled release in three ways.

1. If the API is highly insoluble then attempts to encapsulate it might end up with the API mostly on the outside of the polymer, resulting in "burst" kinetics when placed into a test solution.
2. If the API is highly insoluble but is compounded carefully to be very small particles inside the polymer, its release will follow "Higuchi" kinetics which are linear in square root time.

3. If the API is nicely soluble then its release will follow Fickian diffusion and also be linear in square root time.

What is usually wanted is "zero order release" which is constant with time. There is no obvious solubility/diffusion profile that can deliver zero order release so devices have to be more complex - with some internal Higuchi or Fickian relatively rapid process blocked by some outer barrier that lets through a constant stream of API. One way to do this is if the API's saturated solubility in the inner few nm of the outer coating is low, exceeded by the concentration of API from the internal system, so that the standard Fickian Flux=SaturatedConcentration/Thickness formula is obeyed.

The above sounds rather obvious. My personal experience with controlled release formulators around the world is that many of them were unaware of one or more of the logical steps behind designing a rational controlled release system. As I wrote in my chapter on PLA, much hopeless work on controlled release from PLA/PLGA could have been prevented via some elementary thinking about solubility science.

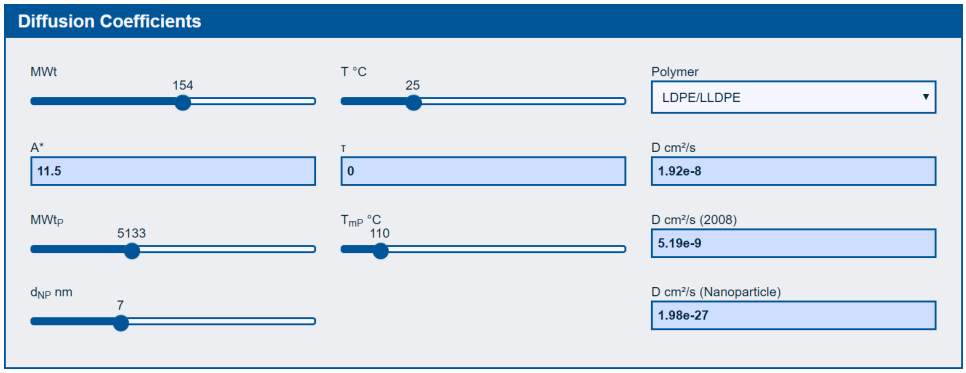
The most practical way to think through polymer/API solubility issues is via HSP, combined with ideal solubility. HSP are becoming increasingly popular[79] in controlled-release publications because they provide a rational approach that is far more successful than the classic "hydrophilic/hydrophobic" approach. Given the restricted number of polymers that are allowed to be used for (human) controlled release, any improvements to solubility/compatibility need to come from safe additives/plasticisers. These will also alter the diffusion coefficients. So a combination of solubility and diffusion science is needed for those who aim for rational design.

9.2.5 Obtaining the diffusion parameters

With the apps (or commercial programs able to handle more complex issues such as multilayers) it is trivial to perform diffusion calculations provided you know the two or three parameters that matter. The parameters are the saturated concentration in the top few nm, the diffusion coefficient at near-zero concentration and the coefficient at the saturated concentration if this is high enough to cause a big concentration-dependent change. How can you determine these values? There are many sophisticated ways which you can find in the literature. My preferred ways are by far the simplest.

To find the diffusion coefficient of a molecule of a known MWt in a known polymer at a chosen temperature, just look it up in the Diffusion Coefficient app

79 An attendee at the 2017 Controlled Release Society conference told me that HSP (and HSPiP) approach was "taken for granted" among those using polymers.



App 9-4 https://www.stevenabbott.co.uk/practical-solubility/diff-cdepend.php

The estimates have the great value that they are "official", in that they are the de-facto standard for estimations of permeation in food-stuffs[80]. The regulators acknowledged that it was impractical for everyone to measure diffusion coefficients for every permeant, and that allowing everyone to choose their own estimation scheme would not be a good idea. Piringer and others got enough data from enough permeants, backed up by enough good science to come up with these schemes. The nanoparticle estimates in the app are backed up by experienced researchers, though gaining experimental verification is rather hard.

To find the diffusion coefficient *and* the saturated concentration, take a thin sample of your material, weigh it, then dunk it into the test liquid and take it out from time to time to weigh it. When the weight reaches a plateau you immediately can calculate the saturated concentration. To find the diffusion coefficient and, if the saturated concentration is high, the enhanced coefficient, simply play with the app till the calculated uptake numbers match the experimental values. In the early stages you can eyeball the values, in the more refined stages you can hover the mouse over the graph in the apps to read out exact values. This sounds rather crude, but it is surprisingly quick and effective. Of course you can write your own diffusion engine in Excel (it is not too hard) and use Excel's Optimizer to find the perfect values. Why bother when the app method works surprisingly well?

The only problem is the "thin sample". Make it too thin and the absorption is so fast and the weight gain so small that the errors will be high. Make it too thick (remember, half-time goes as thickness^2) and the experiment will take far too long. My advice is to make a few samples, say, 25, 50 and 100µm thick, test them all at the same time but ignore data that is too fast and give up if data is too slow. It is very little extra work and the extra (albeit less certain) data act as a check on the values from the chosen thickness. The same consideration applies to the "from time to time" phrase. Try every 30min then every hour then ... adjusting the time to the next sample via intuition and common sense. If you start the experiment first thing on a Thursday morning then after your end-of-day

80 The app contains the relevant literature citations

sample you get a nice 12 hr sample time before coming in on Friday morning and a whole weekend's worth for certainty about the plateau level.

9.3 Diffusion conclusion

The simple laws of diffusion are easily grasped but it is not entirely trivial to implement them. With the availability of apps to do the hard work, there is now no excuse for not using diffusion science in the many areas where it is highly relevant. When the science of diffusion is coupled with your favourite solubility tool such as HSP or COSMO-RS, the combination is very powerful for addressing a wide range of issues.

10 The new language of solubility

When I started writing this book, I had no idea what this last chapter would be, and it came as a big surprise to me that I wanted it to be about language. Why might you be interested in such a chapter? Because if you are framing your questions in the wrong language the chances are high that your analysis is automatically misdirected down routes that have reliably failed to give usable answers to real world problems. Asking the right question in the right language is the first step to finding usable answers.

The background to this is that the more I explored new (for me) areas of solubility, the more dismayed I became by what is out there. I would start reading about a new field and get confused. So what's new? New fields are often confusing for the outsider. Yet more often than not I realised that the confusion was *within* the field and not just by outsider's ignorance.

This chapter brings together a number of frustrations from other chapters and attempt to create a more rational language for talking about solubility issues. My contention is that a more careful use of more meaningful language will allow us to pinpoint the issues that need to be resolved rather than (as in many of the areas discussed here) spending decades getting absolutely nowhere.

10.1 Too many schemes

In the first chapter I selected a very few schemes for proper discussion within the book: HSP, COSMO-RS, DLVO, Ideal Solubility and KB. Those who prefer other schemes might well object, and I didn't deny that some schemes such as UNIFAC are really useful in specific circumstances; my concern here is with the general formulator needing to understand a wide range of solubility behaviours. As I started to read areas outside my own comfort zone I expected to discover new and exciting theories that I might want to add. The exact opposite happened. It is clear that too many bright academics are spending too much time devising their own specific schemes to address a specific topic. This is a massive waste of resource and does solubility science, and those academics, no good. The more schemes we have, the less useful each of them becomes, as they take the community away from solving the issues of the really important schemes.

This first became clear when I was doing the research (with Shimizu) for a chapter[81] on scCO2 entrainers. Researchers would not only regularly come up with new schemes for addressing issues of scCO2 solubility, they would come up with papers comparing, say, 5 such schemes to find which one was best. In reality the schemes would often not show much difference, when tested against

81 Steven Abbott and Seishi Shimizu, *Understanding entrainer effects in scCO2*, in Thomas M Attard and Andrew J Hunt (Editors), *Supercritical and Other High-Pressure Solvent Systems*; Green Chemistry Series, Royal Society of Chemistry, 2018, in press

some standard, but usually arbitrary, list of scCO2 solubilities. This isn't how good science is supposed to work. "Fishing expeditions" are OK if there is no other choice. But it is far preferable to have a hypothesis and create a set of data that deliberately challenge the hypothesis. It became clear to us that if the scCO2 community focussed on just two "good enough" equations which linked nicely to a proper KB approach (which, in turn, derived its data from relatively standard scCO2 experimental procedures) then there was a clear hypothesis to explore and a focussed high-throughput effort using well-chosen solutes and entrainers could confirm or refute the approach.

The point is important. A few months of hypothesis-based scCO2 experimentation by a small team would tell the community more than years of ad hoc "stamp collecting" experiments and "fishing expedition" analyses.

The same principle applies to that poster-child of smart soft matter, PNIPAM. There are countless ad hoc attempts to explain what is going on, including the use of three χ parameters or complex thermodynamic schemes where even the authors complain about the complexity of the formulae. These complex schemes have *never* stood a chance of being adopted and, because of the language problems discussed below, have generated near-zero insight or predictive capacity. A hypothesis-led approach says: "Clearly there are complex Polymer-Water, Water-Water, Polymer-Polymer, Polymer-X, Water-X, X-X interactions going on. Our first duty is to get values for those effects via an approach (KB) that very naturally obtains them for modest experimental and computational effort." When everyone can agree on the key effects (for PNIPAM, contrary to many schemes, Water-Water, Water-X and X-X are irrelevant) then it becomes possible to think through the tricky and subtle effects with greater clarity.

A beautiful example of the convergence (albeit accidental) between theory and experiment has been that of Hofmeister and proteins. KB showed that nearly everything commonly believed and used as explanations (water structure, chao/kosmotropes...) was wrong and that specific interactions, readily measured via routine KB data analysis, were the key. KB, of course, could not say what those specific interactions were, but smart NMR and FT-IR experiments, focussed on the same hypothesis, could do this nicely. The accidental element is that the experimentalists reached their conclusions by a different route, which is fine. The point is that the more we can get into the habit of using theories that provide unambiguous language and combining them, via specific hypotheses, to clever experimental techniques, the faster progress we will all make.

Out of all this comes a proposal to reduce the excess numbers of schemes that will never make a significant impact. Here is the proposal, followed by some relevant bullet points.

The whole of solubility science can be handled via "mean field" approaches except for a few key areas which can only be handled by specific "potential of mean force" interactions.

- The mean field approach means, in practice, HSP or COSMO-RS.
- The potential of mean field approach can be disentangled at the *general* level via KB and can be elucidated at the *specific* level with targeted experimental techniques such as NMR or FT-IR and (if possible) MD simulations linked to KB.
- Although KB is not vital for mean field approaches, it should become the default language because it provides useful insights (e.g. translating broad-brush activity coefficients into a sense of how much "clustering" this entails) along with alerts that mean field is breaking down.
- If the solubility community focussed on these two elements (and stopped stamp collecting and fishing expeditions), progress on understanding and predicting solubility phenomena would be much more rapid.

Before exploring the implications of the proposal we have some other language issues to deal with.

10.2 What does "solubility" mean?

It had never occurred to me that the word "solubility" posed any problems. I just took it that something was soluble or insoluble. End of story. The truth is that solubility is a word that describes a large range of phenomena, from classic "yes/no" solubility of crystalline solids to phenomena such as spinodals (e.g. where a polymer is scarcely soluble in a solvent but the solvent is quite soluble in the polymer) and to the world of dispersions/colloids/nanoparticles where it is hard to decide what is "soluble" and what is "dispersed".

Words can shape our experimental world. Someone who works on "dispersions" might never want to use Hansen Solubility Parameters, because the word dispersion instinctively creates the idea of something that is not soluble. Yet the evidence is overwhelming that HSP are hugely useful for solvent-based dispersions.

Similarly, someone might be happy to use Fluctuation Theory of *Solutions* (KB) for solutions of proteins, starches or DNA, but would not choose to use it for *dispersions* of *nanoparticles*. Yet I cannot find any fundamental difference between solutions and dispersions, and the Gibbs Phase Rule approach of Shimizu demonstrates that there *is* no difference. My scouting of the literature about KB and dispersions found only a few references using it in fairly obscure settings, but they were using it very naturally as though it was obvious.

Then we have spinodals. Such phase separations are common in classic miscibility phenomena and it is not helpful to say that A is insoluble in B when

the separated phases contains plenty of both A and B. Again, as discussed under Flory-Huggins theory, when polymer and solvent are too unalike, the system splits into two phases. One phase contains a low concentration of polymer-in-solvent and the other phase contains a high concentration of solvent-in-polymer. A weirder example is the "co-nonsolvency" of PNIPAM. The polymer is happily soluble in either water or methanol, yet it becomes ... And here's the key point. The *word* "co-nonsolvency" automatically implies that the PNIPAM is insoluble but that's simply not true. If you start with a high MWt polymer at high concentration in water and add the right amount of methanol, the cloudy result, if left for several days separates out into a blob of PNIPAM in water/methanol. A specific example[82] tells us that an initial volume fraction methanol:water:polymer 0.235:0.640:0.010 separates into 0.206:0.647:0.147 and 0.238:0.761:0.001. As you can see, the phase that, from the implied name is "insoluble", is a super-high concentration of polymer in methanol:water. As we shall shortly see, the combination of "insoluble" and "hydrophobic" applied to PNIPAM is a way to introduce double misunderstanding *before one has even started to think about the issues*. The real word that should be used for the PNIPAM/water/methanol case is "co-spinodality" meaning that at either extreme the solution is a single phase and in intermediate mixtures the system is spinodal. A different word for the effect of temperature on PNIPAM in water will be introduced shortly.

10.3 Falsifiability

Although falsifiability is not an absolute criterion for good science, the absence of falsifiability is often a good indication that something is wrong with a theory.

My experience of reading countless solubility papers is that they aren't falsifiable. By stringing together words and phrases that sound meaningful, by fitting the data to some specific theory with plenty of parameters that are specific to that theory, it is possible to weave together a paper that purports to explain a solubility phenomenon, with no possibility of meaningfully applying those words, theories and parameters to some situation which might challenge them.

One of the appealing characteristics of DLVO is that it is easy to find exceptions to it. Examples include what happens when there is somewhat too much steric stabilizer polymer so that it begins to cause either bridging or depletion flocculation, or the case of Hofmeister salts where DLVO breaks down because potentials of mean force, rather than the nice smooth classic DLVO curves, are in control. So DLVO is falsifiable.

A specific falsifiability example in HSP struck me profoundly at the time. Hansen and I were analysing some data from quantum dots stabilised by some smart chemistries. One of the datasets failed any rational HSP analysis so we mentioned to the originator of the quantum dots that either HSP was invalid for

---

82 Na Xue,Xing-Ping Qiu, Vladimir Aseyev, and Françoise M. Winnik, *Nonequilibrium Liquid−Liquid Phase Separation of Poly(N-isopropylacrylamide) in Water/Methanol Mixtures*, Macromolecules 2017, 50, 4446−4453

this sort of material (a distinct possibility), or that sample of quantum dots was contaminated. By applying some high-powered analytical tools they were able to find that the sample was indeed contaminated. A rather simple solubility theory, tested with a bunch of test tubes and a yes/no "soluble or not" test was able to reveal the presence of a nano-scale contaminant. Falsifiability is very powerful.

When I first learned about solubilizers and decided that HSP could not handle them, I tried out COSMO-RS. It quickly became obvious that COSMO-RS could not handle them either. It is a mean field theory and so it cannot handle the local effects that drive hydrotropy. I discussed this failure of COSMO-RS with the community who agreed that the theory indeed was a failure in this domain. In this case, as I had expected, the falsifiability has led to a determination within the COSMO-RS team to be able to tackle such issues.

KB is not falsifiable, but that is only because it is pure, assumption free thermodynamics. What it does do is provide falsifiable hypotheses about causes and effects. You cannot argue (assuming the raw experimental data are adequate) if it says that G_{2u} is large. You *can* argue about root causes of a large G_{2u}. For Hofmeister ions we see that NMR and FT-IR experiments show root causes that match the KB expectations. Perhaps other experiments will provide a different idea of the causes. But whatever happens, there is no need for experiments looking for large G_{11} changes if (as is the case here) G_{11} is totally irrelevant. Any paper that explores G_{11} effects in an attempt to explain what we know to be a G_{2u} effect is falsified.

I won't quote the reference, but a 2017 paper on scCO2 investigated the interactions of ethylene and of ethanol with the scCO2. There's nothing wrong with that. What *is* wrong is that they said that the work was important for understanding how ethanol acts as an entrainer. Because we know that in this case G_{12} (CO2-ethanol) is irrelevant to the entrainer effect, their justification for publishing the paper has been falsified. Indeed, there was nothing falsifiable about the paper. They got some numbers (you can always get some numbers) and offered no hypothesis on how those numbers might provide insights that would be useful to the community. This approach to research is all-too-common and should be regarded as basically unacceptable.

Interestingly, for its first 25 years, KB was a non-falsifiable theory of (justifiably) little interest to anyone. When Ben-Naim developed the inversion technique, it suddenly became capable of generating falsifiable insights.

Now take just about any solubility paper in any area that you know about and see if it contains a falsifiable hypothesis which can be transported to some other aspect of solubility. My experience is that being alert to the (non)existence of falsifiability language built into a paper is an excellent preliminary screen for the value of that paper.

10.4 Hydrophobic

This is another term with so much baggage that it should be abandoned in all serious discussions of solubility issues (though I admit to using the word as a convenient short-hand in some circumstances).

The first level of misuse features the quote I've mentioned twice previously: "X is hydrophobic because it is insoluble in water and therefore we dissolved it in ethanol". What is scary about this quote is not so much its absurdity but the fact that many people would read the quote and find little wrong with it.

The far more serious misuse of the word is as an "explanation" for something: "X happened because of the hydrophobic effect." The world is divided into three types of people when it comes to the "hydrophobic effect". The small minority (and that includes me) has no idea what it means. The majority think they know what it means and accept it as an explanation but would not be able to produce a coherent explanation of what it actually is. And they are of necessity unaware of the other two groups. The third group are those involved in never-ending battles where each participants points out that *their* explanation of the hydrophobic effect is correct and the others are wrong. Entropy and enthalpy are used as if they somehow explained what is going on, despite the fact that the proposed entropic and enthalpic changes are generally large numbers which more-or-less cancel out. And, of course, water structure must appear somewhere in these various schemes, a case of a non-explanation being based on another non-explanation.

A simple example shows why "the hydrophobic effect" is so unhelpful. Take some acetonitrile and heptane and mix them together. They phase separate. We can explain this via the acetonitrile effect; the polar molecules prefer to form aligned groups amongst themselves and the heptane molecules are forced to be amongst themselves. Why stop with acetonitrile? For every X that is immiscible with heptane we explain it via an "X effect". In the end these explanations are a way of saying "heptane is insoluble in X because heptane is insoluble in X", a statement that is both true and unhelpful. For heptane, X includes DMF, DMSO, methanol as well as water. So we have 3 cases of immiscibility with solvents that have *only* polar interactions and 2 cases where H-bonding is undoubtedly involved. As it happens, ethanol is entirely miscible with heptane so H-bonding is nothing special.

There is a famous quote from GS Hartley who was a pioneer of the study of the hydrophobic effect: "*The antipathy of the paraffin chain for water is, however, frequently misunderstood. There is no question of actual repulsion between individual water molecules and paraffin chains, nor is there any very strong attraction of paraffin chains for one another. There is, however, a very strong attraction of water molecules for one another in comparison with which the paraffin-paraffin or paraffin-water attractions are slight*."

This is fine, till we realise that *exactly* the same thing can be said in terms of acetonitrile and paraffins.

All through this book we see that we can measure KBIs which tell us the relative tendencies of the species in question to like or dislike each other. For acetonitrile = 1 and heptane = 2, as we start to add more 2 we find that G_{11} and G_{22} become large and positive and G_{12} becomes large and negative. If you try the KB Binary app, you see that the results are nonsense. The app is too dumb to know about immiscibility, so it generates wild numbers.

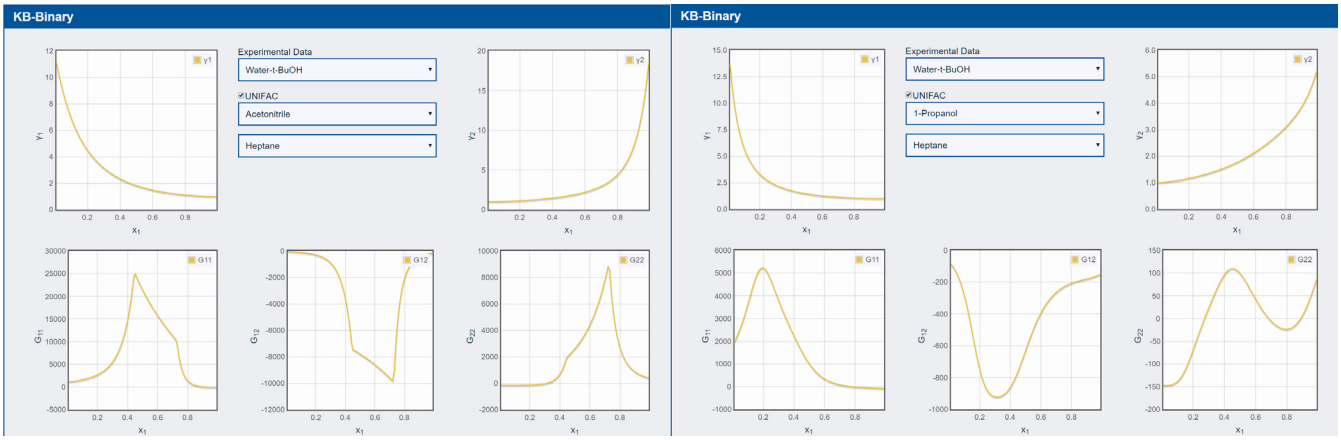


Figure 10-1 On the left: The acetonitrile effect. Or is it just yet another pair of solvents that are not happy together? On the right: The 1-propanol effect?

If, instead of acetonitrile we used 1-propanol we would, not surprisingly, find that G_{11} and G_{22} get very large and positive and G_{12} is very large and negative. 1-propanol and heptane are entirely miscible, yet they show plenty of mutual unhappiness.

There is no difference in *principle* between the acetonitrile and propanol cases and there is no acetonitrile effect which explains something compared to a 1-propanol non-effect. In the case of acetonitrile, the KBI changes happen to be more extreme and, in the end, the theory breaks down when phase separation takes place.

The Hartley point about paraffins not being very self-attracted is more clearly seen in the KBIs for 1-propanol/heptane. G_{11} is *much* larger than G_{22} because the propanols are positively attracted whereas the heptanes are not much bothered with whatever is happening. In water this effect is further amplified in magnitude but not in principle. There is no hydrophobic effect any more than there is an acetonitrile effect.

I promised a double problem with PNIPAM. It is commonly stated that at ~30°C PNIPAM changes from a hydrophilic soluble form into a hydrophobic insoluble form. The delightful "PNIPAM is never hydrophobic" paper by

Pelton[83] shows the double absurdity of this common misrepresentation of what happens at the PNIPAM conformation change temperature. As Pelton says: "Very dilute solutions yield aggregates with a few PNIPAM chains whereas more concentrated solutions yield colloidal sized particles (mesoglobules) or macroscopic precipitates." Those "aggregates" don't seem to me to be "insoluble" polymer. All that has happened is that the polymer has gone from unfolded to folded conformation. This suggests that in the case of the water/methanol effects, instead of co-spinodality we have a co-conformation effect, i.e. the conformation that exists in pure water or methanol changes to a different one in mixed solvents. Both co-spinodality (at high concentrations) and co-conformation (at low concentrations) are more valid and less distracting terms (if we want to have terms) than co-nonsolvency. Such changes of phrase make a huge difference. If the focus is on conformation rather than "solubility" the ways of thinking about the problem are not hijacked by our assumed knowledge about solubility.

More important than the conformation effect is that even when there is phase separation, it is not because PNIPAM is hydrophobic. As Pelton drily says: "At 40°C the water content of the phase separated polymer is 52 wt%, not the hallmark of a hydrophobic material". And, I would add, "not the hallmark of an insoluble material". The use of "insoluble" and "hydrophobic" doubly frame any subsequent discussions which, of necessity, focus on why the hydrophobic effect is rendering the polymer insoluble. Given that both terms are either wrong, misleading or meaningless, the double mis-framing is not helpful if you really want to know what is going on.

10.4.1 Hydrophobic hydration

A fascinating further variation on the hydrophobic theme is the amazingly named "hydrophobic hydration", which is hydration *caused* by hydrophobic groups. It is a phrase casually used throughout many PNIPAM papers as though it *explains* what is going on. The phrase itself is not explained, it is just used as an explanation.

However, one paper[84] is devoted to hydrophobic hydration and PNIPAM. It starts with the general praise of water as an amazingly important solvent and states: *"Besides being the most abundant liquid, the importance of water as a solvent is largely due to its capacity to solvate polar as well as apolar entities. The hydration of the latter is termed hydrophobic hydration; a hydration process that is enthalpically favoured at the cost of being unfavourable to entropy."* I hope that readers are sufficiently alert to notice that the sentence is largely

---

83 Robert Pelton, *Poly(N-isopropylacrylamide) (PNIPAM) is never hydrophobic*, Journal of Colloid and Interface Science 348 (2010) 673–674

84 I. Bischofberger, D. C. E. Calzolari, P. De Los Rios, I. Jelezarov & V. Trappe, *Hydrophobic hydration of poly-N-isopropyl acrylamide: a matter of the mean energetic state of water*, Sci. Rep. 2014, 4, 4377

meaningless. To say that the water "solvates" the apolar entity is merely to say that if the apolar entity is soluble then it is surrounded by water. Well, yes, but what else could the solute be surrounded by? To confirm the suspicion that this is largely meaningless, a few sentences later we have *"Despite the undisputable importance of such processes a general consensus regarding the origin of hydrophobic hydration and hydrophobic interactions has to date not been reached."*

The title of the paper, "... a matter of the mean energetic state of water" is its own refutation. As we have consistently seen, these effects have nothing to do with "mean states" and are all about specific interactions under specific circumstances with ions, urea, sugars or co-solvents such as methanol. Indeed, my final quote from the paper sums up a lot of the reason for this chapter: *"The main reason for this is that hydrophobicity is a multiple facetted problem that depends on the shape and size of the hydrophobic entity as well as on temperature. This complexity has led to extensive investigations and controversial discussions that to some extent prevent the development of simple concepts and guidelines for the wide community of scientists working in other fields like for instance materials and pharmaceutics, where hydrophobic assemblies or the thermosensitivity of amphiphilic polymers are of great interest."*

The problem is not that hydrophobic hydration is difficult to explain, it is that as an explanation it is bankrupt. If you read one of countless papers on IR measurements of the water -OH stretch you find that it changes in hydrophobic environments. Aha! Hydrophobic hydration! But think about it. If the hydrophobe doesn't totally fall out of solution, it is going to be surrounded by water molecules - just as a hydrophobe in acetonitrile is going to be surrounded by acetonitrile molecules. And, yes, those surrounding molecules are going to be a bit different by definition. So what? Acetonitrile's -CN frequency changes according to the solvent environment; how can it *not* change? Because water is supposed to be special, any change in the -OH stretch is seized on as evidence of something deeply significant, yet we never talk about acetonitrilophobic acetonitrilation. From all the papers I've read it seems to me to be evidence that some of the water is sitting next to a different environment, but we don't need IR to tell us that.

Not only does it *not* help, despite decades of being used as if it did help, the idea of hydrophobic hydration *cannot* help because its seeming meaningfulness stops us from thinking in terms that really can help. In *every* case where I have found KB being used to explore issues commonly described in terms of hydrophobic effects and hydrophobic hydration, I have gained more insight and clarity from the KB than from all the hydrophobic verbiage. There is a simple reason for this. KB allows an assumption-free, rather pedestrian analysis *at the molecular scale* of what is specifically interacting with what. The core problem

with hydrophobic explanations is that they operate at the broad thermodynamic level that is explicitly free from any molecular knowledge.

If we want to know what our molecules are doing, then we must use a molecular explanation and not a broad thermodynamic one. As we saw with the Hofmeister effects, KB tells us with clarity what is going on in principle, and smart experiments can identify the specifics.

10.4.2 Anomalous temperature effects in water

I will acknowledge one (almost[85]) distinctive feature about molecules in water, often referred to as a distinctive feature of hydrophobic hydration. Occasionally[86], the effects of temperature are the opposite of what intuition expects. Solubilities of some hydrophobic and amphiphilic molecules are *higher* at *lower* temperatures (a Lower Critical Solubility Temperature, LCST), with PNIPAM being a rather complex example of such a phenomenon. I preserve that sentence from an early draft. It contains at least two errors. The first is that for many molecules it is not so much that solubilities are higher but that molecules which, from scattering experiments or KB analyses of activity coefficients, are deeply unhappy in water at a lower temperature become somewhat unhappier and phase separate at a higher temperature. The second is that PNIPAM does not become insoluble, it changes conformation and may also separate into phases containing different amounts of water. The correct language is far less misleading than my original sentence.

It will be no surprise to the reader that decades of papers purporting to explain this phenomenon have usually led to, as best I can tell, near-zero useful insight into what is going on. If any of these papers had given a definitive explanation, there would have been no need for all the others. My view, defended below by some recent data, is that there isn't very much going on, which is why no one has been able to provide a definitive (and predictive) understanding. For a non-expert such as myself to say "there's nothing much of great interest" in the context of one of the defining and wondrous features of aqueous solubility requires some justification. Fortunately, the examples quoted in the COSMO-RS chapter prove my point.

A molecule that has had the full panoply of techniques thrown at it is t-BuOH in water. It happens to be miscible with water at all proportions and temperatures but scattering experiments show an increasing mutual dislike at higher temperatures. Fortunately (which is why I have chosen t-BuOH over many

---

85 There are examples (in the Francis reference) of glycerol showing LCST behaviour, e.g. with ethylbenzylamine

86 The *Critical Solution Temperature* compilation by AW Francis, Advances in chemistry series; no. 31, 1961, lists ~300 normal water/solute pairs and ~20 pairs with an LCST

other such molecules) we have a full KB analysis of these effects which I have indicated in the diagram.

2-Butoxyethanol (BE) shows similar (but more extreme) KB behaviour at room temperature and then phase separates as a genuine LCST at 49°C.

The other butanols are so insoluble that there is little meaningful KB data. They become soluble in water at ~120°C so their behaviour is "simple" rather than "special". In fact, BE becomes soluble again at 130°C so it shows both types of behaviour.

Now take trimethylamine N-oxide, TMAO. Structurally it is very similar to t-BuOH, but, repeating the image from the aqueous solubility section in the COSMO-RS chapter we see that it behaves completely differently in water.

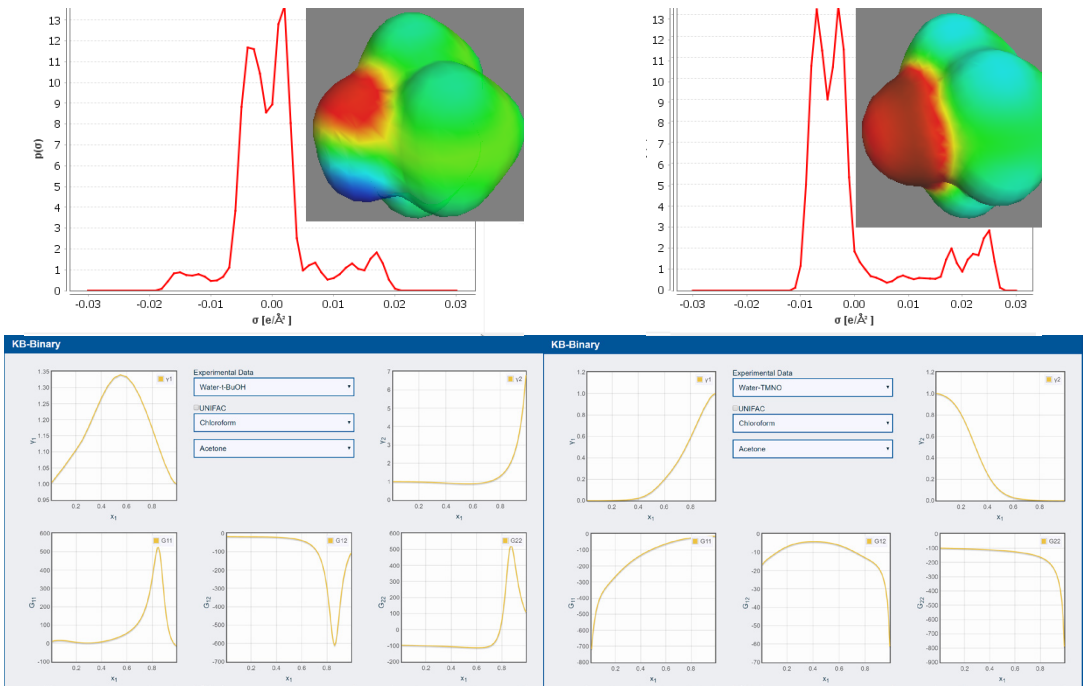
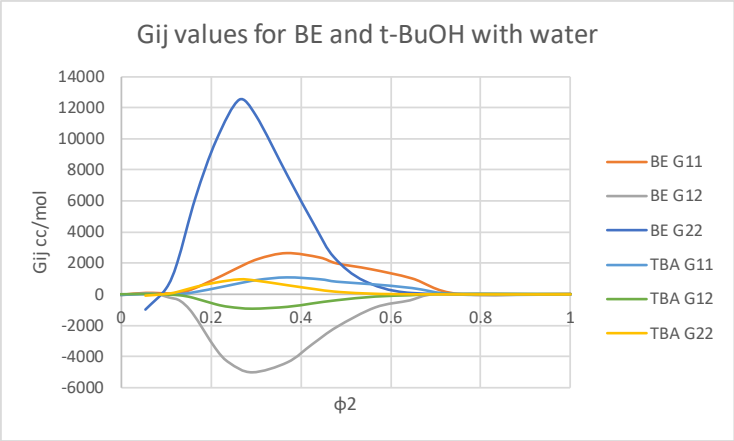


Figure 10-2 Compare and contrast t-BuOH and TMAO with COSMO-RS and KB

Straightforward hydrophobic molecules are basically insoluble so are of no interest. TMAO has everything needed to be hydrophobically interesting because of the big ball of hydrophobic methyls. The reason it is uninteresting is that the NO group forms rather strong H-bonds and that overrides the hydrophobic part. Indeed, an ingenious MD paper[87] takes the structure of t-BuOH and imposes the electronics of TMAO and shows that the relatively modest (less than 2x) increase in H-bond strength is more than enough to tip the balance from "interesting" t-BuOH to "normal" TMAO.

87 Sandip Paul and G. N. Patey, *Why tert-Butyl Alcohol Associates in Aqueous Solution but Trimethylamine-N-oxide Does Not*, J. Phys. Chem. B 2006, 110, 10514-10518



Which leaves us with t-BuOH and BE[88]. At ~20°C the maximum values (in the φ_2 range of 0.2-0.4) of G_{11}, G_{12} and G_{22} for BE are respectively 2000, -5000 and 13000 compared to 800, -900, 1000 for t-BuOH. As a comparison, the entirely dull ethoxyethanol has values of 0, -80, -150. At ~40°C, the t-BuOH values have increased by a few 10's of percent to 1000, -1200, 1100 while the BE values have increase 6 to 8 times, well on their way to phase separation at 49°C. At 55°C the t-BuOH values have merely doubled compared to room temperature.

What does calorimetry have to say about BE? In the range of interest, where the high G_{ij} values take place, the calorimetry (details below) reveals nothing. At a much lower concentration of BE there are some calorimetric effects which arouse the usual stuff about water structure, despite being entirely irrelevant to the phenomenon of interest. It is worth noting that in precisely this low concentration domain there is strong evidence that trace impurities can cause "interesting" phenomena such as 100nm sized agglomerates. I neither know nor care whether the calorimetry is picking up on the artefacts because the significant features are outside the zone of LCST interest. Something about water makes people get wildly excited about stuff even when it is irrelevant.

I mentioned the calorimetry because one of the many "explanations" for LCST is that it is "caused" by an entropic effect. The enthalpy of the water/solute interaction does not change much, so because entropy is always multiplied by T, it has to be, the logic goes, entropy which is causing the effect.

The rather subtle interpretation of "entropy" in terms of the COSMO-RS σ-profile discussed in the context of hexane/water may well turn out to be a useful way to look at explanations for LCST effects. I am taking exception here to the more common use of entropic explanations.

Because as soon as the word entropy is used in the context of water the word "structure" and, possibly, "iceberg" follows. I am happy to report that a number of papers confirm that icebergs are totally irrelevant to these LCST phenomenon.

---

88 For the record, here are the papers used to gather these data: Tadashi Kato, *Kirkwood-Buff Parameters and Correlation Length in Aqueous Solutions of n-Alkoxyethanols*, J. Phys. Chem., 88, 1984, 1248-1252; W Su and Y Koga, *Excess partial molar enthalpies of 2-butoxyethanol and water in 2-butoxyethanol-water mixtures*, Can. J. Chem., 67, 1989, 671-676; Keiko Nishikawa, Hisashi Hayashi, and Takao Iijima, Temperature Dependence of the Concentration Fluctuation, the Kirkwood-Buff Parameters, and the Correlation Length of tert-Butyl Alcohol and Water Mixtures Studied by Small-Angle X-ray Scattering, J. Phys. Chem., 93, 1989, 6559-6565

So what about structure? At this point I must introduce a pair of papers that to my mind destroy the whole edifice of special water effects in terms of LCST.

The first[89] takes the dullest-possible polymer which I will call PLJ, polyLennardJonesium, a string of pseudo atoms with a very simple (Lennard-Jones) van der Waals potential between them. PLJ is then placed into two solvents. The first is a standard molecular dynamics water. The second "molecule" is an arbitrary Lennard-Jones atom arranged to be a "neutral" solvent and to have the same density and coefficient of thermal expansion so that all possibility of explanations involving different interactions from different mean spacings between solvent atoms are removed. Everything is arranged so that both the water and the solvent produce a nice unfolded polymer at room temperature and that the solvent continues to be nice at higher temperatures, while the water causes PLJ to take on a folded conformation.

So we certainly can get the phenomenon of interest, and by stripping away all other complexities the simulation tells us that the *only* difference between the normal solvent and the water is that in water the self-interactions of the polymer monomer units get larger as the temperature increases, thereby tipping the polymer into the folded state. This shows up in (effectively) the RDF of the monomer-monomer interactions (g_{22} which creates G_{22}). The RDF of the monomer-monomer interactions in solvent hardly changes with temperature, while the first peak of the RDF gets higher with higher temperatures in the case of water.

It is all rather subtle, and subtlety is at the heart of a previous paper by the key author[90]. It shows that the hydrophobic behaviour of Lennard-Jones atoms arises from subtle differences familiar to thermodynamicists about constant volume and constant pressure approaches. In turn, the very low thermal pressure coefficient (water is much less compressible than a typical solvent at normal temperatures) means that the distinction between constant volume and constant pressure is more important for water and the unusual temperature dependence of solubility arises because of this divergence.

In both cases, "nothing much is happening" and the balance of outcomes of that nothing much is the equivalent of an LCST. It takes very little for the balance to tip the other way, which is why the behaviour of the majority of solutes in water is "normal".

---

89 I. Hatano,K. Mochizuki,T. Sumi, and K. Koga, *A Hydrophobic Polymer Chain in Water That Undergoes the Coil-to-Globule Transition Near Room Temperature*, J. Phys. Chem. B, 2016, 120, 12127–12134

90 Kenichiro Koga, *Solvation of hydrophobes in water and simple liquids*, Phys. Chem. Chem. Phys., 2011,13,19749–19758

If readers are feeling a little lost by now, stay with me for the second paper[91] that strips away the special nature of the water effect. It involves PNIPAM once more. A very smart Raman set-up allows the unfolded-to-folded transformation to be clearly identified in shifts of CH and OH parts of the spectra. The hydration shell of the folded form is very different from that of the unfolded form, but in my opinion that simply states that the water around anything that changes shape/ size has to be different - it's no big deal. I would be in a minority of one on this point if it were not for the fact that by using a very smart technique they can find that the CH shift happens a long time (in molecular terms) before the hydration shift. The authors therefore concluded that the hydration shell change is *not* the driving force, a sentence I thought I would never see written by respectable academics. Given the famous enthalpy/entropy self-adjustment of water, it is reasonable to expect that the large change in the shell which eventually kicks in is entirely neutral in terms of free energy. The shell, in my view is and always has been a big distraction from the essence which I will now describe.

Here is what I think is going on with solutes in water, and it is nothing very special. For simplicity I will talk about mixtures of two liquids, though the principle equally applies to situation such as PNIPAM. Any two different molecules will have self-self and self-other preferences. If the molecules are not too dissimilar then we see some modest G_{ij} values and that is all. If the molecules are very dissimilar, then there is no significant mixing, and that is all. If they are rather dissimilar then we expect to see large, positive G_{11} and G_{22} values and large negative G_{12} values. If $G_{11}+G_{22}- 2G_{12}$ happens to swing from negative to positive, nasty things will happen to the mixture. The degree of dissimilarity depends on subtle features. TMAO and t-BuOH are similar in structure and % hydrophobic exterior, yet the somewhat stronger H-bonding of TMAO with water means that G_{12} is not very negative. If we go from ethoxyethanol through propoxyethanol to butoxyethanol there is no big surprise that the dissimilarity with water increases. We have no way to know when "insolubility" sets in, but a quick check showed that pentoxyethanol is mostly boringly insoluble, so only butoxyethanol has the supposedly interesting LCST.

Because everything is a delicate balance there is no a priori reason why t-BuOH should have rather modest changes to the G_{ij} values with temperature while butoxyethanol has large changes. An atom here or there in any of the structures could send these delicate balances in any direction. i-butoxyethanol happens to have a lower LCST (25°C) than the n-butoxyethanol; it could easily have gone the other way.

These delicate balances depend on potentials of mean force, which means that the local details around any specific molecule with any specific other molecule can do just about anything, depending on shape, size and charge distributions.

---

91 Kenji Mochizuki, and Dor Ben-Amotz, *Hydration-Shell Transformation of Thermosensitive Aqueous Polymers,* J. Phys. Chem. Lett. 2017, 8, 1360−1364

The PLJ paper (and the paper on LJ solutes) seems to me to be crucial. We have had decades of papers trying to connect specific features of solutes to specific features of water, none of which has provided the sort of clarity we would all love to have. The PLJ paper tells us that the potential of mean force between two Lennard-Jones monomer units and water happens to change rather differently from a more neutral solvent, and that minor difference is enough to tip PLJ over the edge into a folded structure. The difference may be due to the lower thermal pressure coefficient (whatever that is) but it really doesn't matter. Whether a given solvent or polymer shows an LCST is a matter of detail and given the complexity of any real system, there is no point in trying to work out which combination of which details leads to the effect. Think about it. If such a detail was findable in principle then 50 years of good scientists throwing vast arrays of thermodynamics and analytical equipment at the problem would have found it.

All the talk of icebergs and enthalpy and entropy and hydrophobic hydration and chaotropes and kosmotropes and third derivatives of enthalpies has told us precisely nothing. In contrast, a few simple KB analyses of the relevant G_{ij} values gives us a universal language to at least compare and contrast different situations. The language is universal because those skilled in the art can happily convert to scattering, fluctuations or, should they want to, enthalpies and entropies. As we saw in the KB chapters, KB values in binary systems tell us much more than activity coefficients. We will shortly see that things like free energies also tell us very little compared to KB. As soon as a third element is introduced such as urea, sucrose or Hofmeister salts, KB becomes truly transformational because it can say precisely which of the competing G_{ij} are important and which are not. And KB tell us (at least in every example I have looked at) that for three-component systems, specific additive/solute interactions (even when they are as dull as excluded volume) are what influence the system, not general "structure" issues.

10.5 Impossible

The world “impossible” should enter our solubility language, with a precise meaning. There are plenty of solubility issues that are difficult to sort out. If sorting them out is important then the means to do so will become available with better theory or more computing power. The impossible ones are where no sensible amount of theory of computing power will help resolve why effect A happens with solute B while effect C happens with solute D.

In this class I place three types of problems discussed earlier.

The first is the issue of predicting UCST and LCST behaviours of some pairs of solvents. If the calculated curve of free energy with composition is hovering at the zero point over a wide range of composition, the existence of, and the values of, a spinodal simply cannot be calculated reliably and even if the calculation

is correct, the “cause” of B splitting into a spinodal and close neighbour D not splitting will be forever opaque. The entropic and enthalpic terms in each case will be large and balance out on one side or the other.

The second is a polymer such as PNIPAM where the LCST can be shifted by changes in environment such as added salts, solutes, levels of co-monomers, or when the MWt is small and the end-group effects can play a significant role. While the general ideas behind why a polymer shows such a behaviour have been more or less clarified by thought experiments such as the polyLennardJonesium, there is no chance of a useful insight into why any given polymer has an LCST at any given temperature (or has a conventional UCST).

The third is trying to predict interesting phenomena such as the pre-ouzo state. Whether any given system goes straight from “soluble” to “phase separated” or makes the change via a pre-ouzo area once again depends on a subtle balance of effects where the second derivative of chemical potential does or does not follow a certain path. A kJ/mol either way and the system will not show pre-ouzo effects.

It seems to me that someone smarter than me can define impossibility in a mathematically rigorous manner so that we could all agree when something simply isn’t worth trying to resolve. In addition to saving large amounts of academic time spent uselessly on impossible problems, having such a mathematical definition will turn a problem into an opportunity.

10.5.1 Using the impossible

Most of solubility science is “linear”. A bit more X will produce a bit more Y. For a lot of what we do, such safe linearity is highly desirable and our standard solubility tools deal very well with these situations.

The impossible parts of solubility science seem to me to have soaked up large amounts of scientific resource to very little effect. This is because the resource has been focussed on trying to explain the unexplainable rather than on rational ways to control the phenomenon in order to use it more effectively. Papers will often start by saying “We are exploring X because it is potentially really useful” then do nothing that can help us to use the effect. How much better it would be if the papers started by saying “We are exploring X to see how we can better control it so it can be used to achieve Y”. Before I am accused of wanting only “useful” science my argument is against those who study something using its usefulness as an excuse, without doing anything to help it to be useful.

The beauty of these systems is precisely that a small change can produce a large effect; we have amplification which is relatively rare in chemistry. So there are two questions:

1. how can we use the amplification
2. how can we direct it with the minimum of effort.

The answer to the “use” question is not part of this book. The “direction” part is easier. If we have a KB analysis of what is interacting with what, and if, as usual, the answer is that G_{u2} is the main thing that matters, it is relatively easy to know whether we want more or less G_{u2} and we have plenty of examples to give us a clue. With proteins we can choose salts (Hofmeister), small molecules (urea, TMAO) or excluded volume molecules (sugars). With other systems it seems to me that the only reliable solubility forces (excluding conventional signals such as thermal or electrical) for tugging the system in our chosen direction are H-bonds and excluded volume. Because each of these is subject to rational analysis we can make informed judgements of what will kick us one way or another over the fundamentally unknowable borderline. When a conventional thermal signal is used, the question becomes how can we control the environment to shift the temperature and/or shift the sharpness of the response. Shifting the temperature is relatively straightforward. For PNIPAM, the main effects are a lowering of the transition temperature and these seem to be generalised excluded volume effects of no great interest. Changing the shape of the response would require a deeper understanding of the mathematics of the response curve. My grasp of the mathematics/geometry of response curves is far too poor to make any positive suggestions of how to go about changing the response rationally. If anyone who has the requisite theory would like me to bring theory to life in an app-able format then I would be delighted to try.

Although I say that the “direction” part is easier, this is only the case if the basic theory is widely known. Reading the extensive “use” literature can be dispiriting because although the potential uses are often fascinating, it is clear that the general unavailability of usable theories has made it hard for the practitioners to develop rational strategies to improve their control over their systems. With a change of focus away from trying to explain the impossible to a way of using, say, KB to help control such systems the situation can change very much for the better.

Again using PNIPAM as a specific example, the comment that most additives lower the transition temperature by "effects of no great interest" needs to be explained. To reduce the transition temperature by 4°C requires 0.4M NaF, 0.7M TMAO and 4M urea. The simplest explanation is that these are just excluded volume effects. For proteins, TMAO and urea are often seen as working in the opposite direction (indeed, TMAO is often added 1:2 to urea to specifically counter urea's effects, via significant TMAO-urea H-bonds), so the fact that they work in the same direction mean that neither is having a significant interaction with the only interesting groups in PNIPAM, the -C=O and -NH groups. The point which I find significant is that I have not found any paper showing a strong, specific increase of the LCST, meaning either than no one has tried to find such an interaction or that with this polymer there is no possibility of strong control

via H-bonding. It is interesting to think of what properties would be required of a controlling agent. TMAO and urea have very different H-bonding capabilities, with TMAO capable of forming *much* stronger H-bonds. Yet we know that TMAO forms very strong H-bonds with water and seems only to have excluded volume effects with proteins, so its ability to interact with PNIPAM is overtaken by its interactions with water. In terms of the classic KB equation, this G_{21} interaction overwhelms any G_{u2} possibilities. So what molecule *could* be added that would show strong H-bonding effects with PNIPAM and keep it in extended conformation? The problem is that to make a strong 6-membered ring H-bonding with the amide group requires a small dipolar group such as, well, water. So maybe it is impossible to exert stronger control over PNIPAM.

Which brings us back to the topic of impossibility. I'm now saying that PNIPAM is doubly impossible. No one will ever explain why its LCST is ~30°C, because of one type of impossibility and no one will ever do anything really exciting with it via a change to the LCST because the system is immune to such shifts. Such a claim is intended, in turn, for two purposes. The first is to dissuade researchers from writing yet another paper on PNIPAM that takes us no further. The second is to incite someone to prove that I'm doubly wrong by providing a breakthrough in PNIPAM understanding. I would *far* prefer the second scenario, because refutation is such a great way to make significant progress. What I fear, however, is that the solubility fraternity will carry on with more of the same.

10.6 Thermodynamics, enthalpy and entropy

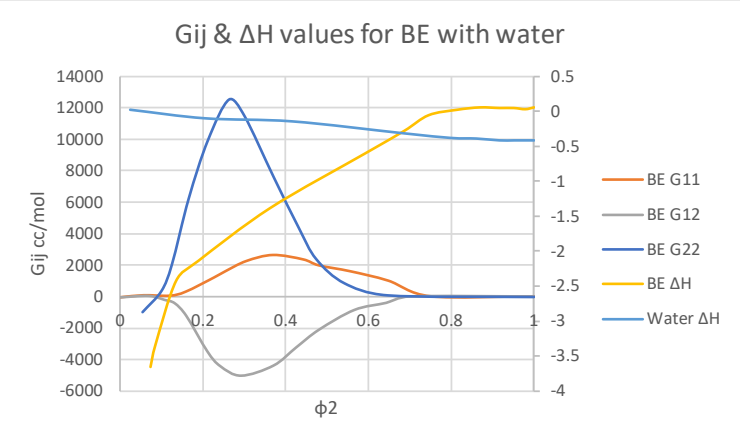
You cannot argue with the laws of thermodynamics. You can, however, argue about the value of applying classical thermodynamics to the questions of solubility that interest us in this book. Let me give an example of what I mean.

Take solutions A and B and solutions C and D. Careful calorimetry of mixing A and B, done at various temperatures gives us a precise idea of the enthalpic and entropic details of what happens when they are mixed. Similar work on C and D tells us that we have very similar enthalpic and entropic effects over the same temperature range. What can we conclude from this about A and B compared to C and D? Nothing! A and B might be two aqueous salt solutions, C and D might be two pure solvents. The high-level thermodynamics are at the same time correct (of course) and meaningless. With the same two pairs and some modest extra amount of work we could do a full KB analysis of the G_{ij} values. Such an analysis would have to know about the specifics of A/B and C/D and would tell us what is happening at the molecular level between A and B and (separately) C and D.

A more remote analogy is to provide the same amount of fuel to two vehicles and to see how far they travel at some constant speed before the fuel runs out. If the distance is about the same we can say that the cars are thermodynamically

the same, even if one is an off-road vehicle and the other is an S-Class Mercedes.

In both cases, the thermodynamics are correct and it is true that classical thermodynamics was and is vital to the development both of solubility theory and of powered transport. The problem is that in both cases, the thermodynamics misses out the features that are truly important to us.



Now look at the thermodynamics of the BE example in the previous section. The same G_{ij} values are there, with the excitement happening at φ_2 ~ 0.26. Superimposed are the "excess partial molar enthalpies". Something interesting is happening to the BE curve when φ_2 ~ 0.1, but whatever that is, it is not relevant to what is happening at the molecular level. Notice, also that nothing at all happens to the water, yet G_{11} *has* to be (and is) interesting if G_{22} is interesting. This is typical of the failure of broad-brush thermodynamics to tell us anything about what is happening at the molecular level. For the same system there also exist entropy fluctuations which, again, peak in the wrong place.

It is not just BE. The whole history of classical thermodynamics applied to solubility issues has produced astonishingly little insight (and plenty of artefacts and side turns) for a very large amount of work. A glance at a set of KB data reveals more than acres of classical thermodynamics ever can.

Moving away, at last!, from aqueous solutions, the classic activity coefficient plots of binary solvent blends reveal amazingly little, as we saw in the introduction to KB. A specific example which solved a big issue for myself and Charles Hansen is the chloroform/acetone system. This system was used to attack Charles during his thesis defence (the Danish "higher PhD" system is like that). The attack had been presented to him in advance and he had to defend himself on the day. Here is the attack. When you mix chloroform and acetone they give off heat because as an H-bond donor/acceptor pair they are mutually attracted. Nothing in HSP can ever produce a net attraction - the best that can be achieved is complete neutrality. Therefore, the claim from Hansen that you can get the properties of a mixed solvent from the weighted average of the two components *must* be invalid for chloroform/acetone. Charles made some predictions on the assumption that they behaved normally, and the predictions were correct. So this "special" system did not behave in any exceptional manner.

How is it possible for a system which defies the key assumptions of HSP to behave normally in practice? I knew of no answer until I stumbled across a PhD thesis (Elizabeth Ploetz at Kansas State) containing some KBIs for chloroform/ acetone. They seemed to indicate that very little of interest was happening, which seemed counter-intuitive. I then ran the binary app and found that there really is nothing much happening.

The graphs are auto-scaling so at first it looks as though there are some dramatic changes. But the estimate of the number of extra molecules around each other is approximately 0.1 of a molecule more than expected, or that at in the areas around 0.5 mole fraction, there was 0.01 extra mole fraction, i.e. 2% more than expected if there are no interactions. So a 50:50 blend is as average as Hansen's assumptions had said it would be.

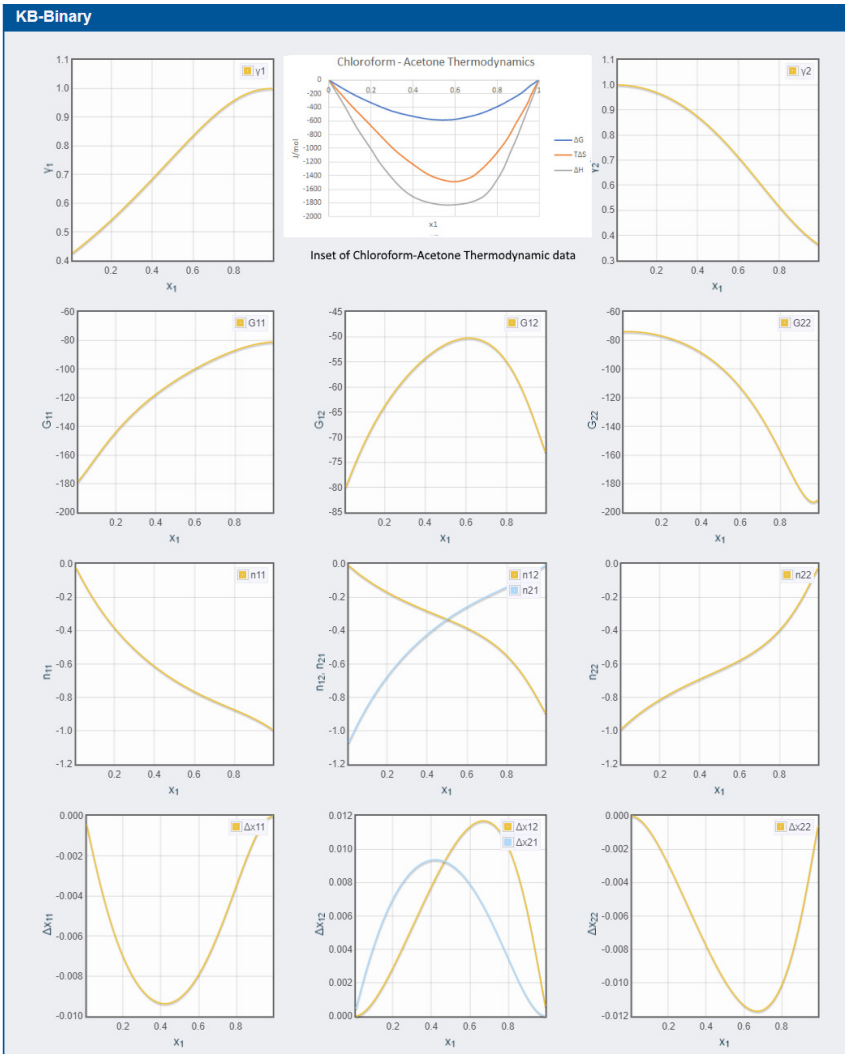


Figure 10-3 KBI, excess numbers and excess mole fraction for chloroform/acetone, with the thermodynamics superimposed

Given that the activity coefficients are rather large and negative (the sign of mutual interactions), and given that the mixing really is exothermic, why are the *molecular* consequences (which is our area of key interest) so minimal? The answer is one that I still find difficult to grasp, even though I know it to be true.

KB is about derivatives of activity coefficients. Large KBI go with large changes in activity coefficients with concentration. What is significant about chloroform/ acetone is that the activity coefficients change gently across the whole range. So the derivatives are never large, so the G_{ij} values are never large, so the excess numbers or excess concentrations are never large.

The thermodynamics of the system are shown as an inset. The ΔG is not very large because, as is so often the case, the ΔH and TΔS terms almost cancel out. The y-axis values are in the 1kJ/mol range, i.e. comparable to RT so in this case the thermodynamics tell us that nothing much is happening. But the KBI tell us in molecular detail how little is happening and that is so much more insightful.

The basic equipment required by the thermodynamicists is not very exciting: calorimeters, densitometers, vapour pressure devices and osmometers. The normal analyses of the outputs of these devices provide astonishingly little useful data in terms of a formulator's ability to understand what is going on at the molecular level. Yet I am expending a considerable amount of my own energies and resources to encourage the solubility community to gather data from such equipment on an epic scale. Equipment is so cheap, and robotics so powerful, that we have the opportunity to create big data on a huge variety of whichever solubility systems happen to be of interest. A synthetic chemist who did not provide a routine NMR spectrum would be laughed out of a journal. A solubility scientist who does not have the key basic thermodynamic data from equipment which is a fraction of the cost of an NMR machine should equally be laughed out of a journal.

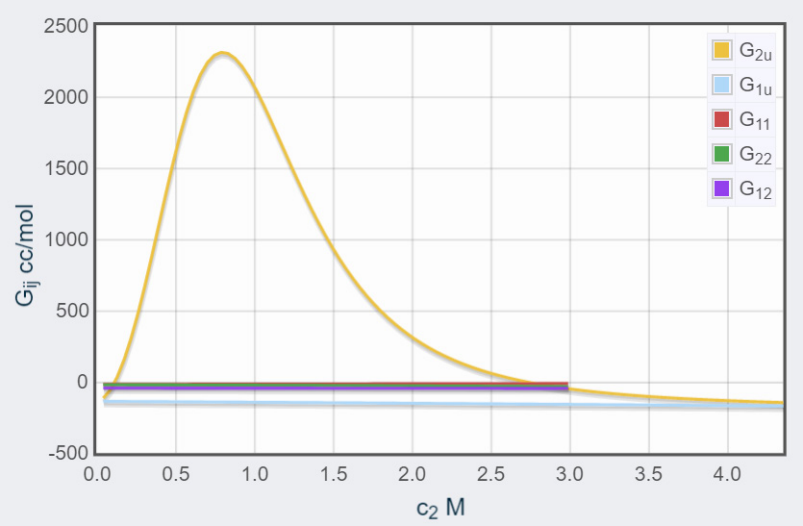
Given that classic enthalpy/entropy data tells us very little, why would we want to gather such data? Because the data can readily be converted into fluctuation theory which, at constant temperatures gives us the molecular level KBIs (concentration-concentration fluctuations) and at variable temperatures gives us concentration-energy and energy-energy fluctuations which are a relatively little-used aspect of fluctuation theory but which, I predict, will take on increasing prominence.

If every paper on interesting solubility issues contained, as a matter of routine, the basic G_{ij} values we would all find ourselves far less distracted by myriads of no-hoper theories and be able to use fancier techniques (NMR, FT-IR, scattering ...) to investigate the effects that are really important (such as G_{u2} for hydrotropes) and de-emphasise those that are unimportant (such as G_{11} for hydrotropes).

This brings us to the final "language" issue.

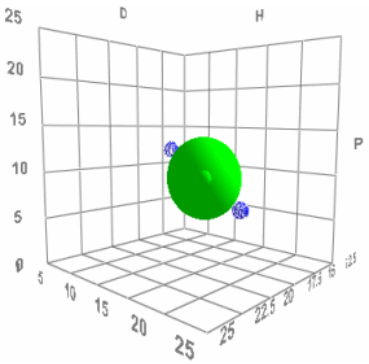
10.7 A visual language

I had finished a consulting task on some fascinating solubility issues for which HSP were well-suited when I was asked, out of interest, if I could explain how solubilizers worked. I had never heard of solubilizers and it didn't take long for me to conclude that I had absolutely no idea how they worked.

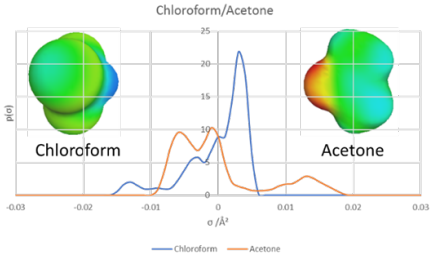


As I hate not being able to answer a question raised in a consultancy, I decided to educate myself in the science of solubilizers and, after considerable effort was none the wiser, with all the standard explanations of water structure and icebergs making no sense to me. Then, after some adventures with densitometers and osmometers, the first plots of hydrotrope G_{ij} values were produced with my colleagues Booth and Shimizu. They were cruder than the one shown here (at the time, we hardly knew what we were doing), but still my jaw dropped. With one graph I understood more about hydrotropes than from all my previous studies, and simultaneously I realised that the graph was indicating that I had so much more to learn. I could not, for example, get my head around the fact that the effect which produced the solubilization (large G_{u2}) became *smaller* (above 0.8M) even when the solubility of the solute was still increasing. You can check the data for yourself in the https://www.stevenabbott.co.uk/practical-solubility/kb-hydrotropes.php app, choosing the BuAc-Urea example. In hindsight the explanation is obvious - at higher hydrotrope concentrations there is so much hydrotrope on average in the solution it does not require an *above*-average amount to keep the solute in solution.

The power of KB comes not just because it is a precise numerical thermodynamic science, but also from the fact that graphical representations of the KBI convey so much information in such a compact form.

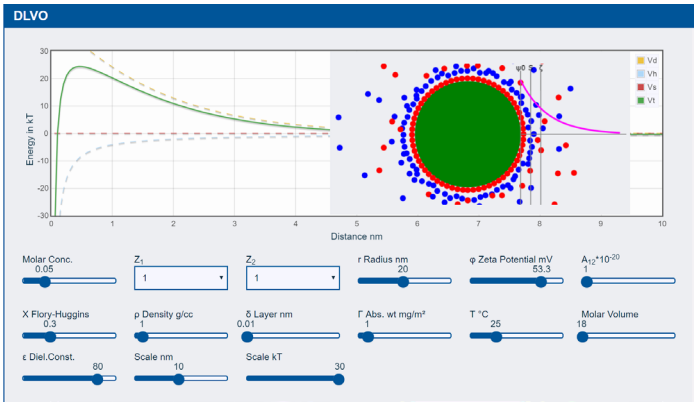


Similarly, one can say that a (bad) solvent with a large HSP distance from a solute can become a good solvent when combined with another (bad) solvent that is also distant from the solute, provided that their vectors cancel out. It is so much more powerful to show this with a simple graphic. I have used this diagram many times around the world and invariably it transforms a "yeh, yeh" into an "oh, yes!". An intellectual idea becomes a usable problem-solving technique.



And the discussion about the thermodynamics of chloroform-acetone becomes so much more alive, and the excellent calculations of this system within COSMO-RS becomes so much more insightful. when the key features of this pair of solvents can be grasped visually.

Anything deeper than a shallow dive into DLVO theory becomes a nightmare for people like me. It is overwhelmingly complex and the extra complexity seems to me to add little value.



So the simple DLVO picture seems to achieve 90% of what is needed, as long as we accompany it with the zeta potential picture as a reminder that we don't really know what that potential is. The strength of DLVO as an aid to formulation is that it allows us to play "what if" with the basic values even if we have great uncertainties about what those values might be. It is especially valuable when any reasonable set of values predicts stability and we find instability. That points to other factors (such as depletion flocculation) being significant.

What I conclude from these (and many other examples) is that our way of speaking about solubility needs to incorporate a strong *visual* element. If the key elements of a theory cannot be visualised, the chances are high that it is one of the many hand-waving or hiding-behind-a-deluge-of-numbers theories that have a proven track record of getting us nowhere in terms of understanding.

10.8 The end to Babel

We have grown up in a babel of solubility languages. The most common one is the equivalent of stone-age grunts, using terms like "hydrophilic" or "polar" as a mask for an absence of anything eloquent to say. The most revered one, classical thermodynamics, is the equivalent of Sanskrit, stark, perhaps beautiful, a root language from which so many others developed, but of limited use for the real world. Then there are myriads of local dialects, meaningful for a very few but of no great help in the wider world. Finally there are the big four languages, used widely, and with plenty of polyglots who can move easily between them. There are also some equivalents of a Google Translate that can allow those not so skilled in one of the big areas to at least grasp the basics of what is going on.

The key elements of the three true solubility languages (i.e. excluding DLVO and ideal solubility) are, to varying degrees:

- A focus on molecular thermodynamics.
- A strong link between the visual and numerical elements.
- A wide range of applicability with a good mix of prediction and understanding.
- Easy access to the key parameters with transportability of parameters from one case to another (unlike QSARs) and a steadily increasing number of such parameters available to the formulation community.
- Good falsifiability. Failures are clear and are used as a rich source of inspiration to better understand the system or the theory.
- A refusal to accept that water is "special" and an avoidance of the convoluted language generated by the aura of being special.

I predict that in the longer term we will converge on just two languages. KB will subsume DLVO and COSMO-RS will mostly displace HSP. The two languages will offer complementary explanations in areas where they comfortably overlap, while providing distinct features in other areas.

There will also be a meta-language which will underpin both languages. In writing this book my respect for MD has greatly increased, partly because KB is such a natural way of extracting meaning from an MD simulation and partly because MD is such a natural way to search for explanations of KB effects. There is a lot of truly bad MD out there, just as there is lots of bad HSP, DLVO and (I assume) COSMO-RS. Against the criteria above for a true solubility language, MD scores impressively well.

I did not expect to end this book with a chapter on language. Having written the chapter I realise that it is something I have been wanting to say for many years but for which I didn't have the language. So much of the near-useless solubility "science" out there is due to the wrong questions being asked in the wrong language, amongst a babel of languages. The aim of solubility science should be to enhance our ability to formulate and to gain control over what happens within solutions. Once we get into the habit of routinely phrasing our solubility questions in one of the great solubility languages we will make much more rapid progress in achieving that aim. Our solubility science will work in principle *and* in practice.

11 Crystallization

When I wrote the original 10 chapters, I'd occasionally looked at crystallization science to see if I could say anything useful or helpful. Each time I looked, I decided that there was little to be said beyond the obvious.

But the scientific life is always unpredictable and I found myself having to have opinions on crystallization. This, in turn, meant that I had to dig deeper into the literature and get to know some of the key ideas in the field. Although many of us need stuff to crystallize for various reasons, and although crystal understanding is important in fields as different as explosives and chocolate[92], it is the pharma industry that has required (and funded) the greatest effort to bring some order to the chaos of crystallization, including the problems of polymorphs - different crystalline forms of the same material with different MPts and solubilities.

For the first part of the chapter I'm interested in creating a "How To" guide. Not so much "How to get the perfect crystals" because that's still an impossibility, but "How to do the least work to get the most understanding". What we've learned in the previous chapters makes much of the "How to" rather straightforward and logical via a step-by-step process using standard solubility tools.

The rest of the chapter is then an explanation of why we need to shift over to an approach based on "clusters" - a change that will make crystallization far more understandable in the long term. It is not a coincidence that the idea of clusters fits in nicely to the ideas of Kirkwood-Buff and fluctuation theory that run through the rest of the book.

Because this is a solubility book, we'll not be looking at crystallization from the melt. That's a fascinating topic, and important for chocolate, but not one that makes sense to cover in this book.

I'm also going to exclude tricks that crystallize via some clever acid-base interaction. By this I mean that to crystallize A is to produce crystals of A. Crystals of salt AB are a perfectly valid way to get A into a solid form, but I will consider that to be crystallization of AB. This declaration that A and AB are different crystallization problems makes the logic of what follows simpler to follow.

Solvent of crystallization is another complexity I will mostly exclude. There *may* be times you want your product to contain an equivalent amount of a solvent molecule, and for molecules crystallized from water there can be advantages to having the hydrated form - the crystal isn't going to be so affected by humidity

---

92 The most stable crystal polymorph of cocoa butter, VI, is "waxy", dull and not "melt in the mouth" while the next most stable, V, is what we all prefer, being glossy, hard and melting in just the right way.

(though there are cross-over temperatures where the hydrated form becomes less stable). But in general, when we crystallize it's to get the pure molecule.

I've also reluctantly decided to skip co-crystals. These have been much hyped, but my understanding is that they are less used than the hype suggests. Why would you dilute your drug with ~50% of something else? And does the extra complexity of screening for co-crystals followed by optimizing for their crystallization and avoiding the multiple complexities of polymorphs really beat solving the original crystallization problem in the first place? Apparently not.

Crystallization requires supersaturation. From a single solvent, this can be obtained by cooling or evaporation. Or we might choose to add an anti-solvent to a near-saturated solution. In thermodynamic terms there is no real difference. In practical terms there can be huge differences. If you need to cool your solvent to 0°C, the kinetics of supersaturated molecules at that temperature will be very different (less thermal motion and also higher viscosity) from those of the same molecules in the same solvent at 30°C supersaturated by evaporation, which will in turn be different from the environment created by adding some anti-solvent molecules. In terms of real-world crystallizers, getting even cooling, even evaporation or even anti-solvent mixing are major challenges - along with the challenge of guiding the degree of supersaturation to be where you want it as crystals fall out.

For the purposes of this chapter, where we have more than enough kinetic complications, we'll assume that single-solvent crystallization takes place via cooling, with the alternative being anti-solvents. However, we will see that a single solvent might be a smart solvent blend acting as a mimic for a solvent with desirable crystallization capabilities but which has a malign toxicity or environmental reputation. The details of how you practically control the process once crystals have started to form (guiding the degree of supersaturation mentioned above), is a topic we'll ignore. Although modelling suggests "3rd power cooling", which means very slow initial cooling post supersaturation, followed by strong cooling towards the end, the logic is based on idealised linear solubility curves which aren't generally the case. And these days, with modern on-line monitors, it's probably best to use a feedback control system to ensure a constant supersaturation.

If the reader senses frustration in what lies ahead, it's not just me. Read just about any paper from Prof Terry Threlfall[93], one of the "greats" of real-world crystallization and you will find exasperation with crystallization itself and with the fact that much effort has gone in to approaches that could never result in useful help for those struggling with a difficult crystallization problem.

---

93 For example: Michael B. Hursthouse, L. Susanne Huth, and Terence L. Threlfall, *Why Do Organic Compounds Crystallise Well or Badly or Ever so Slowly? Why Is Crystallisation Nevertheless Such a Good Purification Technique?*, Organic Process Research & Development 2009, 13, 1231–1240

So here is my step-by-step guide to crystallization from a solubility science point of view. Once we've done the "solubility" bit we'll take on the bigger challenge which is making sense out of how crystals nucleate and grow. Spoiler alert, most of what is written about this ("Crystal Nucleation Theory" CNT) is nonsense so we need to find a more coherent way to think about the topic. "Nonsense" might seem a harsh word, but when there is general agreement (usually after pages and pages describing the theory) that its predictions are wrong by a factor of, say, 10^{10}, that seems to me to be a good definition of nonsense. There are indeed cases where standard CNT works well - it's just that they are at the far extreme of a continuum that describes how crystallization really works.

One final introductory item. Many years ago when I was struggling to understand adhesion science, a wise colleague said "Adhesion is a property of the system". That remark changed my whole approach to how I teach, write about or use adhesion science. It was rather late in the writing of this chapter that I realised that life would be much easier if we all recognised that "Crystallization is a property of the system".

11.1 Step 1: Could I *ever* crystallize this molecule at scale?

Let's imagine an ideal solvent for the molecule. For the moment, let's put aside considerations of cost, toxicity, greenness of the solvent - we just want a solvent where solvent and solute are equally happy being in their own or the other's company, i.e. their activity coefficients are 1 and there is no thermodynamic penalty within the solution. We might dream of some super-molecular interactions, acid/base, donor/acceptor that really drag the solute into the solvent, but such effects are either super-strong, in which case we are making a salt, or add complications such as producing crystals containing "solvent of crystallization" which are less soluble than the pure crystal - for the obvious reason that if they were *more* soluble, it would be the pure crystal which would settle out.

Those who remember Chapter 3 will recall a discussion about ideal solubility. If you know the solute's MPt, its enthalpy of fusion and its Δ heat capacity (ΔCp) as a virtual supercooled liquid you might be able to calculate the maximum solubility within an ideal solvent. It's an amusing pastime to read books on crystallization science because they all have earnest discussions about ideal solubility theory that go on for several pages, yet they all end with the same conclusion: that something like the Yalkowsky estimate based on MPt is the best balance of insight versus hard work. Here's what was said back in Chapter 3:

Fortunately, the careful analysis of this unhappy situation by Yalkowsky tells us that the uncertainties in our knowledge of the two uncertain parameters is generally large enough that we can forget about them and use the delightfully simple alternative for which we only need to know T_m:

Equ. 11-1

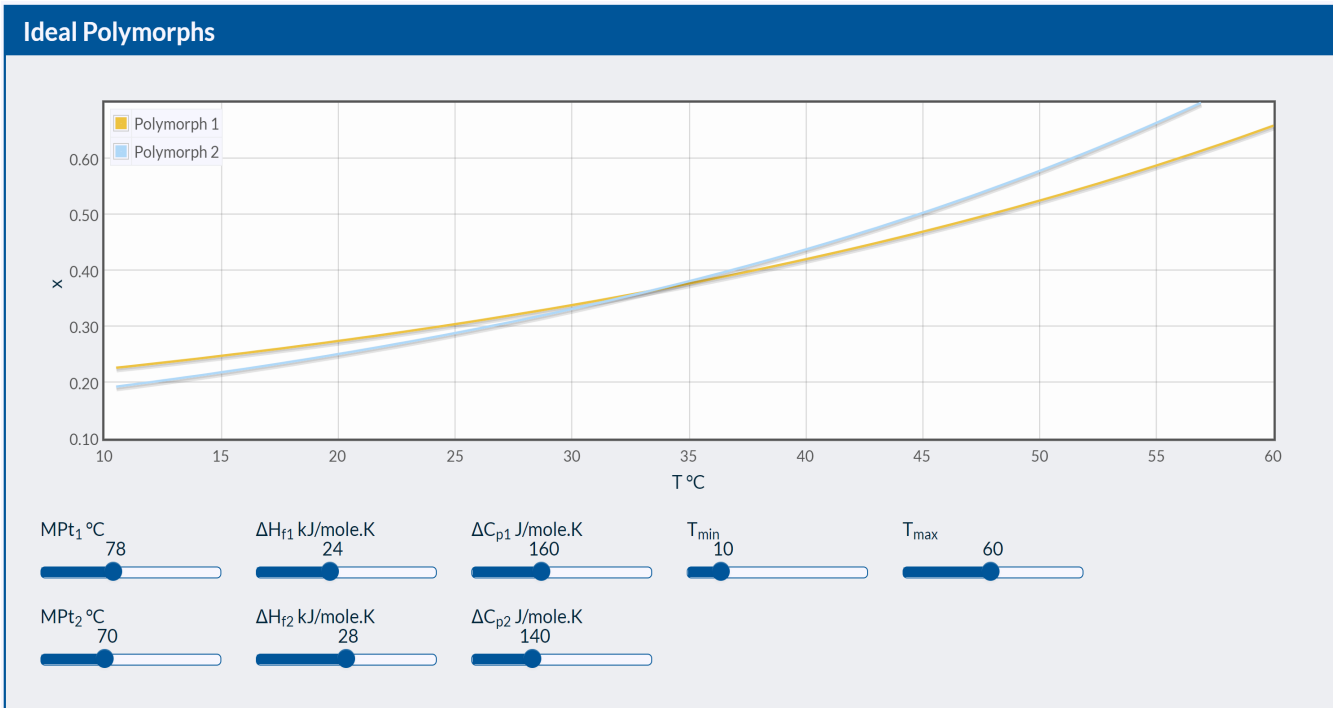
$$\ln(x) = -0.023(T_m - T)$$

However, when we get into polymorphs, things are more serious. Their solubility curves may remain almost parallel or they may cross. The behaviour requires the full theory, again copied from Chapter 3:

Equ. 11-2

$$R\ln(x) = \Delta H_F\left(\frac{1}{T_m} - \frac{1}{T}\right) + \Delta C_p\left(\frac{T_m}{T} - \ln\left(\frac{T_m}{T}\right) - 1\right)$$

To see the equation in action and how it applies to polymorphs, the code from the original ideal solubility app has been re-purposed, and the axes flipped. We will look into the meaning of the graph later - for the moment just accept that such calculations are possible, giving plots of solubility (in this case mole fraction x) versus temperature:



App 11-1 https://www.stevenabbott.co.uk/practical-solubility/Ideal-Polymorphs.php

There are two key points behind using the simple idea of ideal solubility.

First, starting (probably) with just the Yalkowsky estimate, if your MPt is very high (maybe so high that the molecule degrades before you measure it) then the chances of you getting a lot of it into solution at any temperature much less than, say, 100°C are small, so you'd better get used to choosing solvents with high boiling points. That's the bad news, and especially bad if your solute is temperature-sensitive. The good news is that at room temperature, or going down to sensible low temperatures, you will get a large solubility

difference which makes it logically possible to choose a single-solvent "cooling crystallization" route.

If, on the other hand, the MPt is low, the ideal solubility is going to be reasonably high across a practical temperature range, so cooling crystallization will not be efficient. For these cases it might be better to use the anti-solvent crystallization method. So just by knowing the MPt of your molecule, you already have a good idea of the key issues ahead.

The second point is that polymorph behaviour, in particular whether you have a crossing point or not, depends on two parameters that are (in principle) easy to find - MPt and ΔH_f and one, the virtual ΔCp which is not so easy. In practice, getting reliable MPts and heat capacities from polymorphs can also be tricky, making a bad situation worse. Indeed, my take on a more detailed analysis[94], which requires even more parameters, is that the gain in understanding in terms of polymorphs is negligible, because (as you will find in the app) any behaviour is a balance of obscure parameters over which you have no real control. This probably means that we can just ignore more complex ideal solubility discussions. But at least you have a couple of apps to remind you why you (usually) shouldn't bother.

To take the first step from ideal solubility, you need to find an estimate of the activity coefficient γ in your chosen solvent. A handy plot that captures ideal (you can set γ=1) and non-ideal behaviour, and also plots in mg/ml instead of mole fraction is in the Solubility-T app: https://www.stevenabbott.co.uk/practical-solubility/Solubility-T.php.

Before going to the next step in our thought process, we need to look at the various zones that complicate the apparently simple process of measuring our solubilities.

---

94 J.Th.H. van Eupen et al, *The solubility behaviour and thermodynamic relations of the three forms of Venlafaxine free base*, Int. J.Pharmaceutics, 368 (2009) 146–153

11.1.1 The metastable, and other, zones

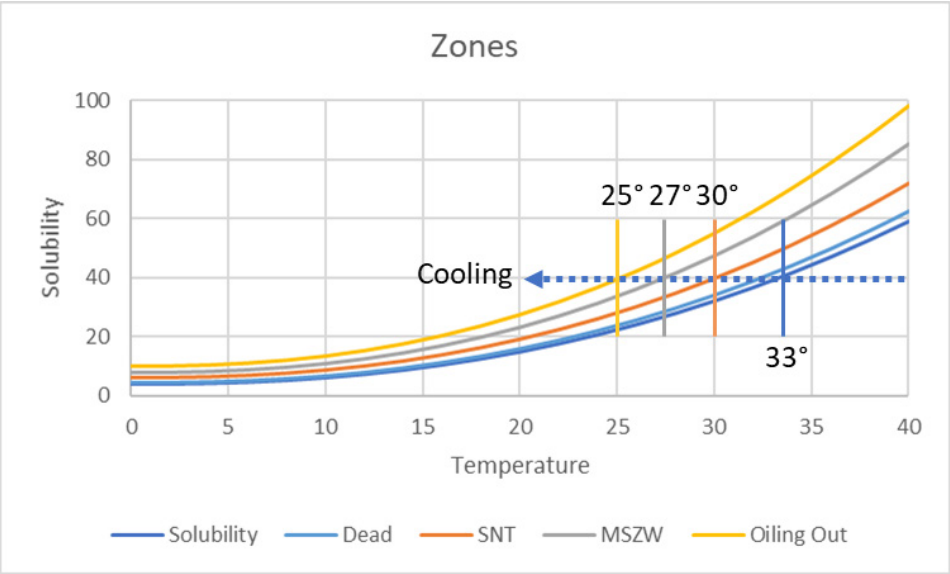


Figure 11-1 Crystallization science likes to talk about zones

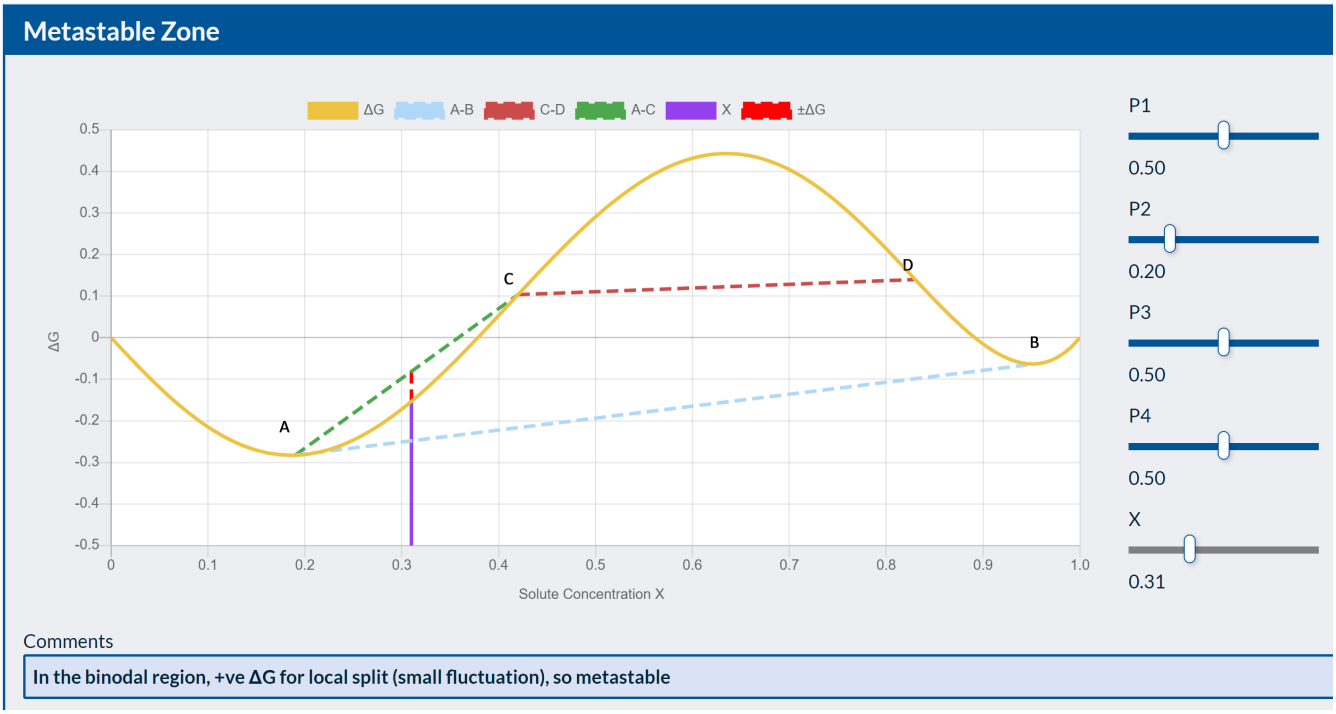
If you cool this solution with a concentration of 40 (units unspecified) from a temperature of 40° down to ~33° then waited a long time you will find solute coming out of solution - you've reached the solubility limit. A subtle point is that there is often a "dead zone" where even seeded crystal growth doesn't happen[95]. At practical cooling rates you might find that nothing happens till you reach ~27°, when crystals are rather likely to form quickly. The difference between the 33° and 27° temperatures is the MSZW, metastable zone width. If you cooled much faster you might find that at ~25° you get drops of oil, you've passed into the oiling out zone. If you are doing a seeded crystallization and want to get some extra seeds (to keep particle sizes small) then you need to be warmer than 27° (uncontrolled crystallization) yet cooler than ~30° which is the SNT secondary nucleation threshold. Obviously if you start at a higher concentration then each of the zone temperatures is higher, but the general trends are the same. These zones, and their acronyms, are important aspects of any crystallization system. Your specific system may have all 4 zones clearly distinguishable or may just have a MSZW so narrow, and crystallization so direct, that in practice there is only a single solubility curve. Naturally the shapes and sizes of the zones will depend on solute-solvent interactions and finding ways to think of these rationally is important for reaching a rational set of compromises for a practical system.

For the moment we will work with just the solubility curve and the MSZW. Even with this simplification, measurement is frustrating for two reasons. The first is that rates of dissolution, near the solubility limit, can be very low. So if you decide to measure solubility by providing excess solute, and leave the crystals

---

95 The dead zone and the SNT are discussed most helpfully in Terence L. Threlfall and Simon J. Coles, *A perspective on the growth-only zone, the secondary nucleation threshold and crystal size distribution in solution crystallisation*, CrystEngComm, 2016, 18, 369-378

stirring for a few hours, you see that they've not dissolved, so you decide that the solution is saturated, filter, measure the concentration and that's your solubility. You had no way to know that leaving them another few hours would have produced further dissolution (unless you use a smart PVM or FBRM system[96] that can continually monitor the crystal size to see if they are still shrinking). Going the other way, as you supersaturate nothing happens till you go so deep into supersaturation that everything falls out too quickly - you've passed through the metastable zone. Although the borders of the metastable zone are difficult to measure, they aren't arbitrary - they represent when the system moves from binodal to spinodal instability. For those who are interested, the text on the app explains what 's going on.



App 11-2 https://www.stevenabbott.co.uk/practical-solubility/Metastable-Zone.php

Just note, for the moment, that the app describes the sort of binodal/spinodal situation that applies to systems which show liquid-liquid immiscibility. So a solute might decide to separate as a crystal, but it might also decide to separate as a liquid, i.e. oiling out. The important thing to grasp in the latter case is that the two phases are *not* oily solute and solute-free solvent, but the mixtures shown at points A and B in the app - one portion might contain 60:40 solute:solvent and the other portion 10:90 solute:solvent. The proportions *within* each phase are controlled by thermodynamics, the *ratio* of the volumes of the two phases depends on the original proportions of solute and solvent.

After that digression, let's get back to solvents.

---

96 Everyone uses PAT (Process Analytical Technologies) with Process Video Microscopes (PVM) and Focussed Beam Reflectance Measurements (FBRM).

11.2 Step 2: What solvent(s) should I use?

An alternative question might be "What solubility estimation tool should I use?" There the answer is simple: COSMO-RS[97]. As argued throughout the book, although there are plenty of ad-hoc solubility tools, the only one with a power and accuracy good enough for this task is COSMO-RS. That's not to say that it's a perfect tool - solubility is far too complex for that, especially when "non-mean field" effects kick in. But there isn't a better tool and given that often for a new molecule you don't have large quantities of already-purified material for doing lab experiments, it's far preferable to get a good idea of the appropriate solubility options before heading for the lab.

You use a DFT (Density Functional Theory) method to get the basic conductor screening model charge distribution around the molecule in a few likely conformers. That takes time and computing power, but these days it is routine. The example used below was produced on my laptop in about 10 minutes. You then take your list of "solvents I'm willing to try" and get COSMO-RS to calculate the relative solubilities at 20 or 25°C, a process that is effectively instant. Let's unpick both of those statements.

In your specific industry, there will be a set of solvents that are loved (in pharma these are the ICH Class 3), a set that are accepted (still Class 3), a set that are reluctantly accepted (Class 2) and a set that are banned. If you start your screening with only the "loved" solvents, the chances are that they are so bland and boring that you'll not get anywhere near the ideal solubility. If you include banned solvents in your screen you might face some internal opposition. Even so, there's a good rationale for including some of them, for reasons we'll see shortly. In any case, COSMO-RS doesn't get tired, so a longer list is probably a better idea than a shorter one.

The calculation of "relative solubilities" is all that COSMO-RS can manage without extra information. If you feed it a MPt, ΔH_f and ΔCp it will do its own version of an ideal solubility calculation and give you estimates of real solubility. If you happen to have a real solubility measure in any one solvent, you can give that to COSMO-RS as a calibration value and its calculations of other solubilities will be scaled accordingly.

Rather quickly you now have a list of solubilities in a range of solvents. If you believe that you are going to use single-solvent crystallization, then you can probably exclude those with a large solubility at room temperature - you'd have to crystallize at very low temperatures to get an efficient process. Using some judgement, and now focussing on solvents more likely to be acceptable in a

97 I note that I have no commercial interest in the various commercial tools for COSMO-RS calculations. It's simply a fact that it's the only methodology out there with sufficient power and proven capability, within the limits of basic solubility theory.

real-world process, do some calculations of solubility versus temperature over the range of, say, 0-100°C.

Although the temperature dependence of solubility happens to be full of its own complexities, you at least have some idea of which solvents stand a good chance of giving you an acceptable solubility when hot, and low solubility when cold to produce a good balance of *productivity* and *yield*. Productivity is a measure of the amount of solvent needed, yield is the % recovery after crystallization. If you need 100g solvent for 1g of solute, that's not a good productivity, while 100g of solvent for 10g is OK. However, if your 10g of solute in 100g when hot falls only to 5g when cold, that's low (50%) yield with high productivity, while 1g going to 0.1g would be high yield (90%) with low productivity. You can't make final decisions at this stage, but you can get a good feel for the options.

As an example, I have used COSMO-RS[98] to calculate the solubilities of carbamazepine in three solvents. The calculations can be checked against experimental data cited in a paper mentioned later.

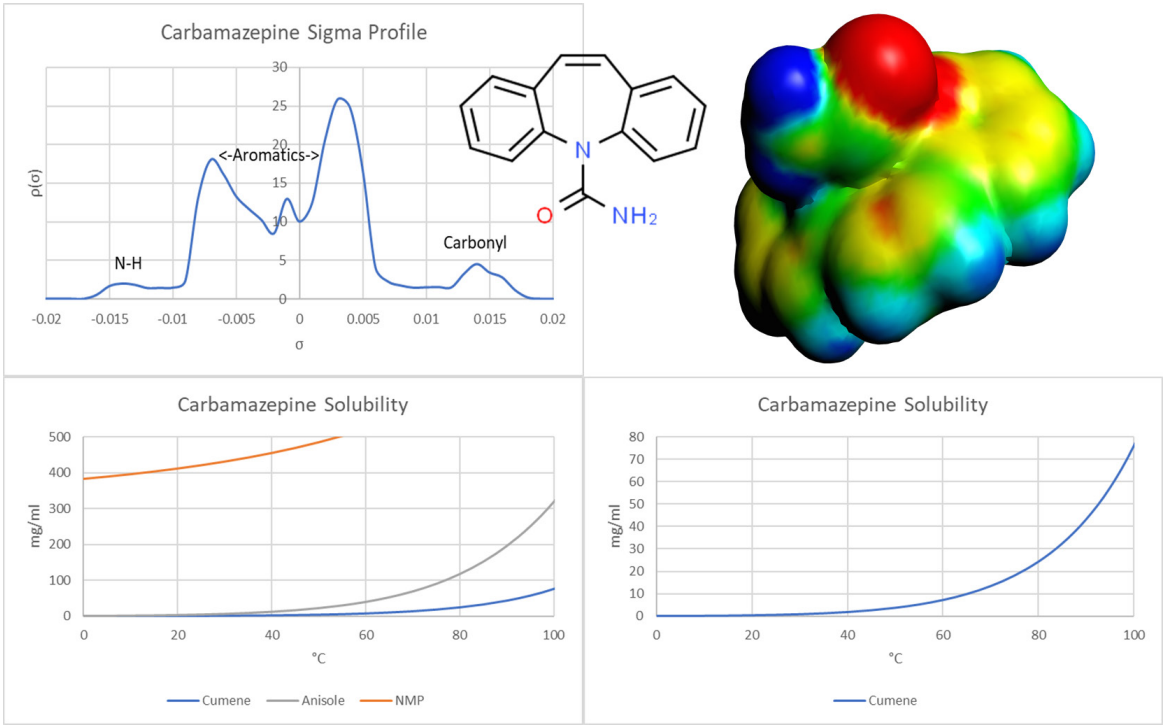


Figure 11-2 COSMO-RS Carbamazepine, sigma profile, sigma surface and predictions of solubility

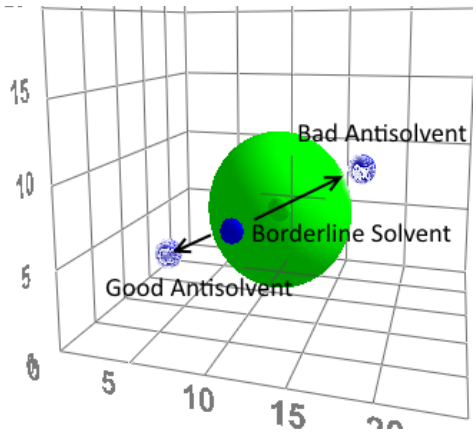
A quick visual check of the 3D sigma surface shows the significant charges on the aromatic rings and these appear in the sigma profile as peaks around ±0.005, then the strong carbonyl charge (red) at ~0.015 and the somewhat weaker N-H (blue) with a peak around -0.015. It doesn't take much imagination

98 For this chapter I have used the COSMO-RS program in the Amsterdam Modelling Suite from SCM. The data are from calculations using default parameters. In more expert hands no doubt accuracy could be improved.

to see that the molecule can interact strongly with NMP, less strongly with anisole and only via the aromatic interactions with cumene. The solubilities in those three solvents are shown, with NMP being far too soluble for sensible cooling crystallization, anisole being marginal and cumene, having both a high BPt and low MPt, having a good balance of yield versus productivity.

As it happens, the relative curves, based on the experimental values of MPt and ΔH_f are a good match for the experimental values, but the calculated absolute curve for cumene is rather too optimistic, about twice the experimental value, at high temperatures. Playing with COSMO-RS it was possible to improve fits by providing a ΔCp value of ~-100J/mole.K. Given that it is experimentally near impossible to determine such a value independently, and that there is no solubility software that can calculate such values meaningfully, we have to accept that the main COSMO-RS errors arise from our poor grasp of ΔCp rather than defects in the software itself.

If it is clear that cooling crystallization won't work out, your solvent list at room temperature gives you a starting point for anti-solvent crystallization. Before starting on a COSMO-RS set of calculations, it is worth looking at a simple HSP idea, repeated from Chapter 3.



The idea that the addition of a bad solvent can make a better solvent has an important negative consequence for those who wish to use anti-solvents to precipitate a solute at a chosen moment, e.g. for crystallization. If, as in the diagram, we suppose that we have a borderline good solvent then we can precipitate the solute with either of the two bad solvents. The "good" anti-solvent would be excellent because just a small amount will take the system outside the sphere. The "bad" anti-solvent would be a disaster because by adding it, the solubility will actually increase (heading to the centre of the sphere) before decreasing and finally precipitating the solute.

In fancy crystallization discussions, the idea of a bad anti-solvent has been described as "anti-solvent synergy". Although COSMO-RS can readily calculate this, it makes sense to have an idea of where in HSP space your solute finds itself so you can then identify which "good-enough" solvents can be rapidly crystallized via an anti-solvent on the correct side of the sphere. You can input the COSMO-RS solubilities into HSPiP and either fit a sphere to those solubilities or establish a good/bad criterion to give a yes/no sphere fit. Either method is fine. Although it might seem odd to use a powerful software tool to create an estimate for a less powerful one, I find an HSP sphere to be a useful navigation tool because I can easily get lost in COSMO-RS space.

With some sensible choice of good solvents and anti-solvents, run COSMO-RS across a range of solvent pairs to find a reasonable balance of efficiency. You want to use the minimum of good solvent and then add the minimum of anti-solvent. This is trickier than it sounds. If you have a really good good solvent, near the centre of the HSP sphere then you have to add a *lot* of anti-solvent. But that increases the total volume of solvent, so the practical decrease in solubility is less than you might like. Going for a less-good good solvent, near the edge of the sphere, and using a really bad anti-solvent might be more effective. Even here there's a problem. In a real crystallizer, getting good mixing is very hard, so too good an anti-solvent might send the supersaturation much too high in the first moments of mixing - so you have to accept a less perfect solvent pair to help with manufacturability.

There's a key problem to be aware of. A really bad anti-solvent might turn out not to be miscible with the good solvent. As discussed in Chapter 3, just before the anti-solvent discussion, even COSMO-RS cannot reliably calculate binary immiscibility, so be prepared for some disappointment in the lab, where, because of the solute you might even have ternary immiscibility issues.

Now we can deal with those unusable solvents (health, safety, cost...) that were included in your virtual screen. If one of them looks especially promising in solubility terms, use HSP to find a blend of nicer solvents which match the HSP of the unusable solvent. Then refine the blend using COSMO-RS. Although a blend is never as attractive as a single solvent, if it gets you into a sweet spot in solubility space, you can probably live with the consequences. And, even better, if the blend seems to be adequate, you can fine-tune the crystallization performance by tweaking the blend. We will see later that this might help with crystallization rates and with some polymorph problems.

11.3 Step 3: Quick checks

Now it's time for the lab. The aim is to check on the basic solubility predictions. If they are right then the serious crystallization work can begin. If they are wrong, well, hopefully they'll be wrong in an interesting manner from which you can learn something helpful.

Test a few reasonably bad solvents to confirm the insolubility, even at high temperature, and a few good solvents to check that things are OK at low temperatures and very good at higher temperatures. If you start your tests with relatively volatile solvents it's easy to remove them ready for the next quick test so that you aren't consuming large amounts of precious raw material. We're not looking for good crystallization performance; we're aiming to confirm or refine our solubility knowledge ready for the serious work.

But you can do a few quick checks on what look, computationally, to be viable choices of solvent(s) and temperature. If you can get a reasonable amount into

solution and via cooling or anti-solvent addition some crystals start to appear (maybe overnight, or with some seeding...) then you are in a good position to start looking for the right crystallization system.

What about high throughput?

The steps so far have been mostly on the computer plus a bit of hands-on work in the lab. Why would we not just miss out all this tedium and give the problem to a robot to sort everything out? The answer, as has been confirmed all too often, is that formulation and solubility space is far too large and robots are far too stupid. A robot used purposefully to gain a lot of information in the context of a reasonable theoretical background is always my choice compared to lots of tedious lab work. But this means getting the basic theory in place, first, before using up the precious resource of a robot that should always been in high demand for other theory-driven optimizations. A few good lab experiments are worth 1000 bad robotic ones. 100 good robotic experiments, based on the theory are then the sensible way forward.

11.4 Step 4: Some real crystallization experiments

It's time to do some crystallization experiments. We need lots of them and we probably still have small amounts of material, so it's a good idea to use the semi-automated systems that can do a lot of experiments in parallel on 1ml samples with good computer control of things like temperature, plus the ability to take videos of the appearance (or otherwise) of crystals.

Because we've done lots of intellectual preparation, we know what we are trying to achieve. We are *not* yet trying to optimize our crystallization. We *are* trying to make the optimization process as painless as possible.

We have a few candidate conditions of solvents, temperatures and anti-solvents. OK, give them a go. Because we know that it takes time for crystals to appear and that there is a balance to be struck between low supersaturation (slow and inefficient) and high supersaturation (fast but poor-quality) we need to know our solubilities and degrees of supersaturation.

Measuring solubilities to high accuracy is very hard[99], doing so over a range of temperatures and levels of anti-solvent is mind-numbingly tedious and takes more time than most of us can justify - especially if it turns out that solvent X gives poor crystallization whatever happens. We have already seen that there are multiple problems. For example, how can you be certain that you've allowed enough time for crystals to dissolve to be able to decide whether a given level of solubility is achievable? Or how long do you wait before deciding that the

99 Everyone agrees that it's worth the effort to get a precise solubility curve for your production system. That's at a stage much later than the one being discussed here.

absence of crystals means that you really are still above saturation? No one has good answers to these sorts of question.

The pragmatic answer (at least for cooling crystallization) is to use an automated system to do multiple temperature cycles, using whatever seems sensible (turbidity, video, ...) to judge "dissolved" versus "not dissolved". With some pragmatism, an adequate cooling curve can be obtained, suitably smoothed to remove the statistical glitches.

Now, starting with a fairly high level of supersaturation, check what happens. If nothing happens, or stuff oils out, well, that's bad luck - the system may not be the one you want. If stuff crystallizes out too rapidly then (a) you get a rough idea of the crystal habit and (b) you can just dilute somewhat and repeat. As this is all statistical, your automated system has to do this a few times before you can be confident that (a) you can reliably get crystals and (b) they are the sort of form you can work with in the long term - you might be unlucky and polymorphs are appearing already to confuse things. Because both oiling out and too-rapid crystallization can arise from crossing the metastable zone too quickly, repeating with a slower cooling rate might get you out of trouble.

Unless you have vast resources, even with these automated systems, scanning a few solvent systems is a serious undertaking. So use your knowledge of the solvents and solute to be smart with what you test. As was stressed earlier, one of the strengths of COSMO-RS is that you get nice pictures of the charge distribution. If from your theoretical scouting experiments you can find two solvents that give you the solubility profile you want, but one solvent is likely to have strong donor-acceptor interactions with one part of the molecule and another solvent is either more neutral or has interactions with a different part of the molecule, those are the two solvents you need to try first. If one crystallizes nicely and the other is a problem, that gives you a working hypothesis that, say, *this* sort of solute-solvent interaction is helpful while *that* sort is not. The hypothesis may later be refuted, but that's fine. Refutation is also new knowledge.

Let's see what the previous paragraph means. Unless you have access to some amazing crystallization prediction software, the only thing you can quickly work out is that either the molecule has such a strong crystal form preference that it will always come out in one form, or it will be influenced, for better or worse, by the solvent. If your first tests are with two solvents that both have similar donor-acceptor interactions with the solute, you'll not be able to tell. But by choosing two very different solvents you learn very quickly how much, or little, the solvent affects the crystal form. If the interacting solvent gives the right form and the non-interacting the wrong form, or vice-versa, you are well on the way to understanding your system and having a chance of a rational optimization process. To put it more specifically, if you have a choice between comparing ethanol with butanol or ethanol with MIBK, choose the latter pair because

although butanol *will* be different from ethanol, MIBK will be different in a more interesting way.

If one solvent gives needles and the other gives platelets, then the crystallization rule that the crystal form is controlled by the slowest-growing face gives you a hypothesis to work with. A strong solvent interaction with one specific group might strongly inhibit growth in the direction of that group. There is, of necessity, much hand-waving about such effects, but we have to work with what we've got. Any single solvent effect will be interpretable in multiple ways, but an intelligent choice of solvents can start to give you workable hypotheses. They might later be refuted, but, again, a refutation is valuable new knowledge and progress in understanding. At this stage use whatever solvents allow you to explore hypotheses, even solvents that you know you will never use in the final process. DMF, chloroform, NMP and γ-butyrolactone will not feature in your final process, but they are each in a unique part of solubility space, with interestingly different functional groups, and can provide precious information.

Too much of anything is a bad thing. Too strong a donor/acceptor interaction between solute and solvent will give you solvent of crystallization which is generally not preferred. If the non-interacting solvent gives entirely the wrong form, then use COSMO-RS (or sort some optional solvents by HSP Distance and find a close one) to find an interacting solvent that is similar to the one that interacts too strongly, though with somewhat less donor/acceptor potential. Maybe that will do the trick. Now you can also see the potential benefit of a smart solvent blend to replace a good solvent that happens to be on the "never use" list. If that blend contains different functionalities, tweaking it might be enough to give you the right level of donor/acceptor without being too much in either direction.

What is a "strong" interaction between solvent and solute? The answer is fuzzy. If the solute has hydrogen-bonding capability then interaction with an alcohol will always be "strong", either as donor or as acceptor. Yet there are papers arguing, reasonably, that the alcohol's *self*-interaction might be stronger, so the net interaction with the solute will be weak. If you use a solvent such as acetone which is a not-very-strong hydrogen bond acceptor, it might still have a strong interaction with a donor bond on the solute because acetone has no self-bonding capability. Again you have to use COSMO-RS and common sense to work out what is going on, then tweak the system in a rational manner to see if it heads in your desired direction.

None of these thoughts are new. The idea of using the solvent (rather than specific crystal engineering additives) in a rational manner to steer crystallization in a desired direction has been around for a long time.[100] The problem seems

---

100 For example, the classic paper: Nicholas Blagden and Roger J. Davey, *Polymorph Selection: Challenges for the Future?*, Crystal Growth & Design, 3, 2003, 873-885

to be one of lack of holistic understanding of those solubility aspects that have to be juggled at the same time, and, more recently, fear of using test amounts of solvents that will never be used in a final process but which can provide so much early information to support or reject plausible hypotheses.

There are no guarantees with any of this, except one. If you try to high throughput your way to a conclusion, you will consume vast resources. A process driven by (imperfect) hypotheses is likely to give you more understanding, and a better outcome for less effort. For example, if you have good IR, Raman or NMR set-ups integrated into your workflow you can choose to get spectra of everything and hope that some AI will dive in and find what you need. A better use of such a resource is to choose to scan for specific peak shifts for specific situations and relate those to your working hypotheses. You might miss a spectrum or two, but at least you won't drown in a sea of data.

Ways we might be more efficient at this via better/deeper science appear near the end of the chapter once we've seen all the ways that don't/can't help.

11.5 Step 5: Sorting out the mess of polymorphs

If you aren't from pharma then you might simply be happy to have an easy crystallization system and not worry if it can swap from one crystal form to another. If you've heard fairy stories about stable polymorphs that suddenly become impossible to attain after "infection" by a different polymorph, well, it's a very real effect[101] but you might still be lucky and if you unexpectedly get a new polymorph your end users might not be too worried. But even in non-pharma areas, polymorphs can be unacceptable. For example, some explosives have polymorphs with different densities, so you'd (usually) want the highest density, and some have polymorphs with different susceptibilities to shock, so you'd (usually) want the least sensitive.

If you are from pharma then polymorphs can be a matter of life and death, in terms of profitability and, in the extreme, in the patient getting too much or too little of the required dose in a given time, depending on the polymorphic form and its kinetic and thermodynamic behaviour. The profitability part has arisen in cases where sales of a popular drug (e.g. ritonavir) had to be shut down for an extended time when a different polymorph appeared two years after the start of mass production, and where (such as the famous ranitidine hydrochloride case) vast profits depended on polymorph patentability. It also matters hugely because getting the right polymorph impacts the day-to-day processes of handling crystals (if they come out as needles they can clog a reactor), or making them in tablet form where different crystal forms can have significantly different mechanical properties. It is the impact on the tablet as a whole that can have a

101 Here's a fairly modern review built upon an equally delightful review about disappearing polymorphs, seed crystals etc. This article is Open Access. Dejan-Krešimir Bučar, Robert W Lancaster, and Joel Bernstein, *Disappearing Polymorphs Revisited*, Angew Chem Int Ed Engl. 2015, 54, 6972–6993

big impact on dissolution rates, rather than the rather modest impact on rates or solubilities (typically a factor of 2) from the small changes in MPt, ΔH_f and ΔCp that differentiate polymorphs.

You might think that in the 21st century we'd be able to compute our way out of the polymorph problem. However, we hit a familiar issue. The difference in energy between polymorphs is often less than 4 kJ/mole, when the positive and negative interactions that, inevitably, balance out are in the 100s of kJ/mole. The excellent Facts and Fictions review[102] shows that if you know the polymorphs' crystal structures you can use DFT to get rather good ordering of the energies of different polymorphs, which is impressive. However, what you can't do is create a plausible range of crystal structures for a molecule (many hundreds!) and predict the likelihood of there being polymorphs. The review also, sadly, disproves some well-known rules-of-thumb suggesting which compounds will give more polymorphs. That review also points out that the initial enthusiasm for solving the polymorph problem via vast programs of high throughput robotics, was followed, as is so often the case, by disillusion. There seems no substitute for the hypothesis-driven approach advocated here.

It is the mix of thermodynamics and kinetics in the crystalline form that causes all the problems. In simple terms, each polymorph, I, II, III ... has, in sequence, a lower MPt[103] and "therefore" if you believe ideal solubility, a higher solubility. The most stable, I, with the highest MPt has the lowest solubility. The free energy difference of two polymorphs is simply $RTln(Sol_I/Sol_{II})$, and as this is negative, in the long run, II must turn into I - though whether this conversion can happen in the solid state (enantiotropes) or needs to go via a soluble state (monotropes) is an important issue, explored shortly in the app. If, again, you believe in simple ideal solubility, then the ratio of solubilities does not depend on the temperature, so although you need different temperatures to get the same solubility in different solvents, the free energy difference between polymorphs does not depend on the solvent.

In the 19th century, Ostwald suggested that we might be able to get polymorphs to appear neatly in turn - with crystals coming from the state into which it was easiest to fall, rather than the ultimately most stable state. In reality the polymorph you get can depend on subtle solvent, temperature, time balances plus the ever-present danger of a rogue seed sending your crystals into the wrong polymorph. This assumes that you know you have III, II and I in that thermodynamic order. But you can work for years on a molecule believing it has 3 polymorphs then wake up one morning to learn of a 4th, 5th and more.

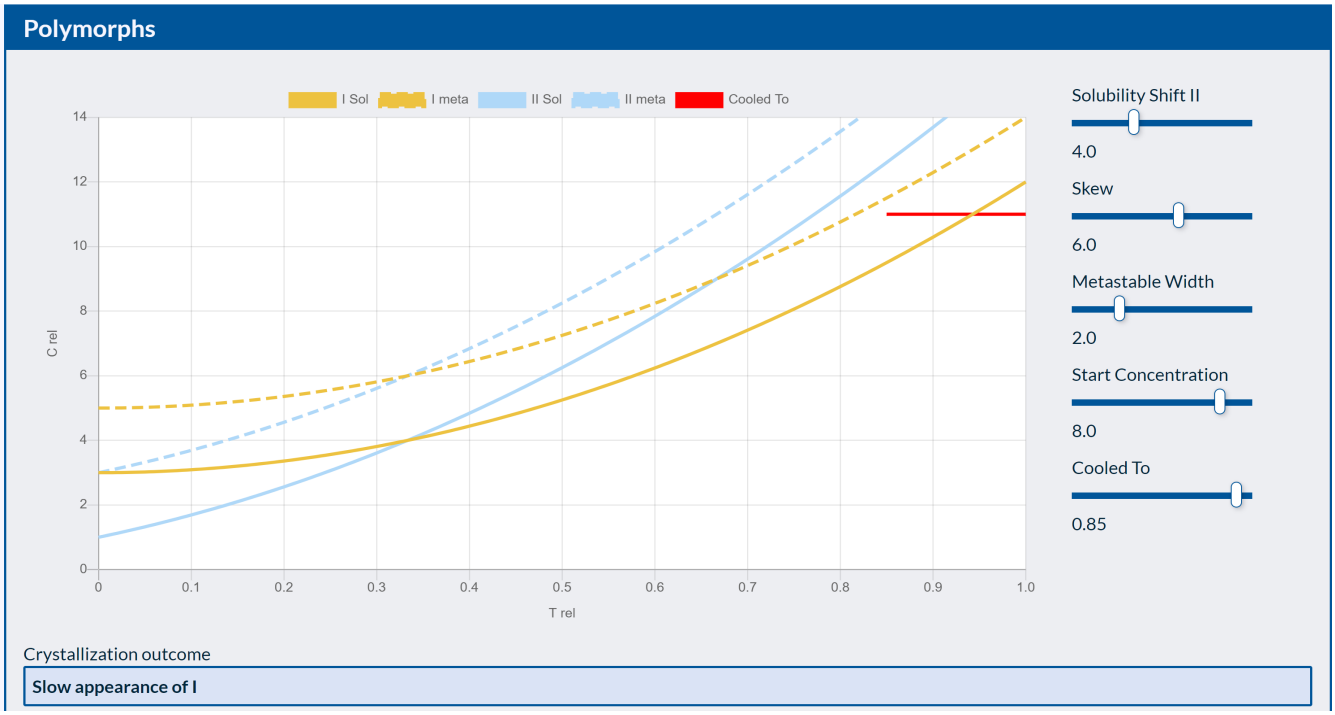
---

102 Aurora J. Cruz-Cabeza, Susan M. Reutzel-Edens and Joel Bernstein, *Facts and fictions about polymorphism*, Chem. Soc. Rev., 2015,44, 8619

103 Unfortunately there are two conventions, I, II, III increasing or I, II, III decreasing MPt. Because of the app I am following Threlfall's preferred convention.

The science and art of determining the number, properties and inter-conversions of polymorphs is a topic too important (the patent aspects alone are a huge topic) to be dealt with lightly, but too complex to be dealt with in this book. Instead the focus is on how the solvent can be used to influence (or not) the outcome. The app is based on a much-cited paper by Threlfall[104] which combines a healthy dose of frustration with some caution that there are at least *some* rules to which we can cling. First, the frustration: *"The author recently undertook to crystallise 20 well-known pharmaceutical polymorphic pairs, using apparently well-described recipes, often from well-respected groups, but failed to obtain the expected outcome in respect of the form obtained in 10 of these cases."* And now the rules.

Because solvents can have such a strong influence on polymorphs, there is a danger (and I'm guilty of it) of missing conditions where the solvent can have no influence and which, therefore, are precious zones of comparative order within the chaos. Following Threlfall, we can identify conditions where form I is guaranteed, where it is highly likely, where it is surprisingly likely ... and where form II is guaranteed. As he comments in the paper, describing the situation with just two polymorphs is difficult enough, so don't complain that the app doesn't include the option of a third polymorph.



App 11-3 https://www.stevenabbott.co.uk/practical-solubility/Polymorphs.php

We have two solubility curves of concentration versus temperature, marked with solid lines. Polymorph I has a higher MPt and lower solubility than II but at a certain temperature the two curves cross and II is the lower solubility and the thermodynamically preferred form. The Ideal-Polymorphs app at the start

104 Terry Threlfall, *Crystallisation of Polymorphs: Thermodynamic Insight into the Role of Solvent*, Organic Process Research & Development 2000, 4, 384−390

of this chapter lets you explore the conditions where you will or won't get such crossings. It also reminds us that it is the whole package of MPt, ΔH_f and ΔCp which combine, in obscure ways, to give us the relative behaviours. Simple texts on polymorphs tend to over-emphasise MPt, when the other factors are often more important because MPt values often differ by just a few °C, not enough to explain much of the distinctive behaviour.

As we saw with the Metastable Zone app, cooling below the solubility curve (going from right to left along the red line in the graph) takes you into the metastable zone (gone past the binodal), at the edge of which (dotted lines) (past the spinodal) the crystals will appear quickly. Determining the solubility curve is hard enough, determining the edge of the metastable zone is even harder[105], so don't worry too much about the precise details.

The sliders allow you to change the two sets of curves to mimic a wide variety of relative behaviours of the polymorphs[106], but let's stay with the one in the screenshot. At a chosen starting concentration (11 on the arbitrary scale) you cool down to a chosen point. The app tells you what to expect. In the screenshot you are in the metastable zone so the likely outcome is slow appearance of I. If you cooled a little further, you'd be beyond the metastable zone so you'd have rapid appearance of I. What about cooling even further so you are below the cooling curve of II? It's debatable, but you'd probably still get I.

Now start the cooling at a concentration of 2 on the arbitrary scale. From visual inspection, and as the app shows, you find yourself entirely in the (now) thermodynamically preferred II plus you will have got there without being close to I being insoluble, so II is guaranteed.

One key point of the Threlfall paper is that the zones between the two cooling curves are pretty much guaranteed to give I when above the crossing and II when below. If you change solvent, the absolute positions of the curves in the vertical concentration axis will change, and maybe the slopes, but the logic does not change. Therefore the polymorph in these cases is *not* controlled by the solvent. This is the big point of the paper and a rebuttal to those like me who over-concentrate on solvent effects.

When you start exploring other zones, the app suggests the probable outcome, though in some cases the result will depend on specifics such as rate of cooling, whether or not you add a seed, and the relative widths of metastable zones. Threlfall is *not* saying that the polymorph is always easy to guarantee or always solvent independent. Instead he is saying that the alternative viewpoint, that polymorphs are controlled by solvent-specific interactions can also be wrong. If

---

105 As we saw near the start, you can also find oiling out zones to further complicate things.

106 If the polymorph curves cross then they are enantiotropic - they can interconvert in the solid state; if not, they are monotropic and can only interconvert, from unstable to stable, via solution.

you always want the same polymorph, and if, as his quote shows, it's not easy to repeat what others say is a good (solvent-specific) recipe, then making sure you are in one of the guaranteed zones is a reasonable strategy.

There is another important point that emerges from this. There are many cases where people have thrown 100 solvents and 20 temperatures at a problem, with disappointing results. But unless you know (and how could you!) the solubilities and saturation levels in each of these solvents at each of the temperatures, you have no idea where you might be in potential zones and, therefore, whether you are over/under-saturating giving too much or too little time. This point is raised in a paper[107] that manages to get all three polymorphs to appear under rational circumstances (explicitly agreeing with Threlfall) from a single solvent, chosen thoughtfully. Better a few rational experiments than 100 irrational ones.

The one thing you cannot do is hope that the problem of polymorphs won't appear. Focussing too early on one solvent system is a recipe for disappointment[108]. Everything works well for months then some strange seed appears from nowhere and a new polymorph appears. You've invested all that detailed work and now a lot of it has been trashed. Instead, you need to apply intuition and cunning to get the polymorphs to appear as soon as possible. Because that combination of intuition and cunning is another hypothesis driven approach to crystallization, it gives you a chance to produce the polymorph you want, when you want, either by being in the correct part of polymorph space, as per the app, or by being in a space controlled by the solvent. If you have, say, 4 very different solvent systems, with at least one of them rationally tunable as a blend, with others being neutral, donor-dominated and acceptor-dominated then it's highly likely that any molecule that is happy to form multiple different forms is going to be persuaded by those solvents to appear in at least a few different polymorphic forms.

Because polymorphs have different optimal forming temperatures, you can get them to inter-convert by some cunning manipulation of temperatures, times and bits of co-solvent that entice enough of one form to start to inter-convert to cause the whole lot to shift.

From different solvents you can get seeds to throw into other types of solvent. If the seed is ignored and you get that solvent's natural form then you know that this specific solvent is relatively robust (or that you're in one of Threlfall's safe zones). If it forces the new form to appear then maybe that's the form you want to continue with. There's no guarantee, but what we're trying to do is provide increased chances of success with the minimum amount of hard work.

---

107 A. Getsoian , R.M. Lodaya, A.C. Blackburn, *One-solvent polymorph screen of carbamazepine*, Int. J. Pharmaceutics 348 (2008) 3–9

108 The authors of the One-solvent paper used just one solvent to make a point - they explicitly endorse using multiple solvents when dealing with new molecules.

If the most stable form is the one that you want, then saturating any less stable form in an appropriate solvent and leaving it long enough (the "slurry technique") will result in the desired form. If you use a poor solvent, the saturated amount will be too low to provide a convincing combination of solubility and super-saturation to drive the conversion. If you use a really good solvent, then the kinetics will probably be better, but unless you can later cool the saturated solution sufficiently, your conversion yield will be low. And although thermodynamics always wins, your patience (and the economics) might be against success, especially if the converting solvent happens to like to block the slowest-growing face of the most stable polymorph. In many cases, "long enough" is days or weeks, but there are surely cases where this is months or years.

If you want to find stable forms via anti-solvents then anti-solvent diffusion, the vapour slowly coming from a separate chamber, is slow and controlled.

More normal anti-solvent addition can give you a mixture of thermodynamics and kinetics, with smart choices of anti-solvent functionality allowing you to explore hypotheses of encouraging or blocking potential faces via strong interactions with functionality on the anti-solvent.

If you want to find unstable polymorphs then kinetically driven conditions are required. The most extreme is reverse anti-solvent addition, where you drop your solution into an excess of anti-solvent, remembering to use a borderline good solvent along with a good anti-solvent that doesn't take the solute through the centre of the solubility sphere.

If you try to solve polymorph problems through brute force high throughput, and fail, you might hope for, but be denied, another round of resource-hungry experiments. If you try a smart combination of theory, experience and high throughput, you might still fail, but at least you can blame the molecule and not yourself or your robot.

11.6 Impurities

Threlfall in the review on the problems of growing crystals cites a patent dispute where extensive analysis of impurities in crystals became crucial. The opposing sides found many 10s of impurities, some of which had no plausible source but were undoubtedly present. The presence of impurities is not surprising; the paper gives the calculation that a 1µg crystal, containing 10^{16} molecules will, at 99.9% purity, contain 10^{12} impurity molecules. This then raises two opposing questions:

- Why aren't the impurities better excluded, giving, say, 99.999% purity?
- Why don't the impurities stop crystallization from happening at all?

There's no good answer to either question, though it is undoubtedly the case that some specific impurities can stop some specific crystallizations. Indeed, one of the potential explanations for why certain polymorphs appear only after months of playing around with just a few previous polymorphs might be that there is more material, of higher purity, and a crystal poison, specific to that one polymorph happened to be, but is no longer, present. It really seems that the more difficult problem is why, given how difficult it is to get even the right molecules to grow in the right position, the wrong molecules, even 1 in 10,000 get a chance.

If you are fortunate enough to know what your chief impurity is then this needs its own solubility analysis first via COSMO-RS then a few intelligent tests in solvents relevant to your crystallization. If you are lucky enough to have two acceptable solvents for your solute and the impurity is much more soluble in one than the other, it's likely that you'll get better crystallization and purity of your material in the solvent preferred by the impurity.

The intriguing idea, discussed below, that impurities can be overwhelmed and kicked off the surface by clusters is one of the many reasons that the cluster-based view of crystallization needs to become the default.

To explain that remark about clusters we need to shift from the solubility discussions to an analysis of how crystallization takes place so we have a better chance of solving problems rationally. The single most important step we must take is to move away from what most people consider to be the standard theory, CNT, Crystal Nucleation Theory. Why? Because the theory is wrong and, even worse, diverts us away from a much better way to think through things - which is via the cluster-based view.

11.7 Why we must abandon Crystal Nucleation Theory

Just about every major resource on crystallization devotes much time to CNT. It's the standard text. I'll spend only a little time and provide you with an app to show you what CNT is, and then explain why it is useless for us. There is still some excitement about a supposedly newer version called 2-step CNT but it's only a subset of what's really going on and is not at all new. Although it seems shocking to dismiss a theory that has been around for almost a century, it really isn't of practical use, and generations of academic papers have not reached any conclusions that could not better be reached without it.

11.7.1 The behaviour we'd like

What we want is to be able to bring (via cooling or anti-solvent) the solution to a point of just being supersaturated so that crystals start to form. By being just supersaturated, they don't form too quickly, so they grow with the sort of discrimination that gives us beautiful crystals that exclude all impurities. As the

crystals start to come out, we need to reduce the temperature or add more anti-solvent to keep the system supersaturated. The result is an efficient conversion of a concentrated solution to one so dilute that we haven't lost much material, but where we haven't tried so hard that impurities start to crystallize out as well.

Unfortunately, this is rarely what happens.

11.7.2 The behaviour we get

In real life we supersaturate and nothing happens, so we supersaturate a bit more and still nothing happens. Eventually (we've crossed the metastable zone) everything crashes out as impure, malformed crystals or as an annoying oil.

If we're patient, we hit an optimal supersaturation then wait till crystals start to form fast enough to be useful but slowly enough to form nice crystals. If you have one of those robotic systems that can do 100s of recrystallization experiments on 1ml samples by moving just above then just below supersaturation, you will find that at a given level of supersaturation, the first crystal might appear after 10s in one experiment then 1000s in the next experiment in the same tube. It's an annoying statistical game. If you try the same experiment with 10ml samples you might "only" have to do 10s of experiments, and with a 100ml test, you might get the same result each time.

If you are doing crystallization on an industrial scale, you have to decide on an optimal level of supersaturation and, perhaps, accept a "standing time" which, with or without a seed crystal added, will give you the crystallization you require.

11.7.3 The CNT non-explanation

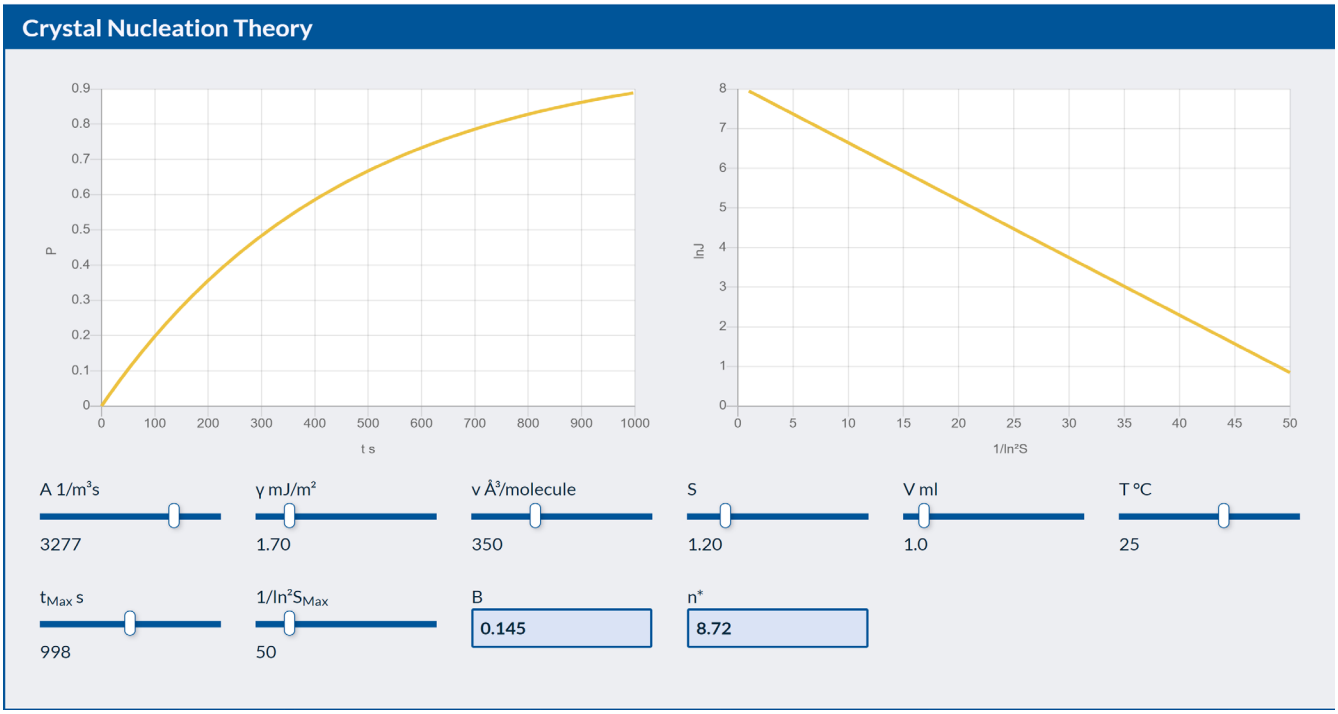
The standard story is that at supersaturation, little sub-nuclei form and un-form spontaneously but when, by chance, they grow to a critical size, the thermodynamics leads these nuclei to continue growing to form regular crystals.

CNT tells us that any molecule has a crystallization rate J which depends on the degree of saturation S[109]. The larger S, the larger J. Because of the random nature of nucleation, to get J you need to do many experiments to find the probability of getting a nucleus (or, rather, a crystal as you can't see a nucleus) at any given time. J is calculated from the probability curve.

Finally, the relationship between J and S tells you about γ the surface energy of the nucleus (J decreases with increasing γ) and n* the critical nucleus size. The app does all the hard work for you:

---

109 This might be a number such as 1.1 which means that the concentration is 10% higher than the saturated value under those conditions. Annoyingly, there are at least 3 measures of supersaturation at concentration c versus the solubility limit c*: c/c*, c-c*, (c-c*)/c*. Each has its uses.



App 11-4 https://www.stevenabbott.co.uk/practical-solubility/CNT.php

There are multiple reasons why all this is bogus. What you observe are (large) proto-crystals, not nuclei, so your experiments never see what they claim the theory is based upon. Worse, if you are seeing "turbidity" are you sure that those are crystalline scattering centres - might they be fluid clusters, discussed below? And what everyone agrees upon is that if you use super-filtered, pure solutions and super-clean tubes you will be waiting a very long time to see crystals - in other words, CNT is often based on observing proto-crystals formed round junk in the system. And the supposed γ values seldom make much sense - and in any case what are you supposed to do with them? As the app shows, you can extract other kinetic and thermodynamic parameters from the data and attempt to find correlations with all sorts of factors (such as MWt) - all to no avail because the basic theory is wrong, not least because nucleation depends so strongly on junk rather than pure thermodynamics.

To emphasise the disconnect from reality, if you write out the equations with many more parameters you can use reasonable values of those parameters and estimate rates that are wrong by 10 orders of magnitude. That's not very impressive.

There's one more relevant fact that I really like. It was, when I first came across it, so surprising that I didn't believe it, and you could sense how nervous the authors were in proposing it[110]. With some molecules you can get a fine balance between two polymorphic crystal forms such that it's roughly 50:50 which form appears on cooling. However, if you gently redissolve the crystals at a few

110 Khalid Hussain, Gunnar Thorsen, Dick Malthe-Sørenssen, *Nucleation and metastability in crystallization of vanillin and ethyl vanillin*, Chemical Engineering Science 56 (2001) 2295}2304

degrees above the saturation temperature, pause and re-cool, you get more of whichever crystal form had previously appeared - in other words there is "memory" across the solubility boundary - with the memory fading the longer you pause before re-cooling. That makes no sense in terms of standard CNT and fits perfectly with the cluster ideas that make CNT obsolete.

Despite decades of measuring J values, γ values and n* values for many solutes in many solvents, the total number of useful insights from all this work seems to me to be 0.

Or, perhaps to be kinder, they confirm that:

1. The more supersaturated, at a higher concentration, the faster, on average, stuff crystallizes.
2. Solvents that can interact strongly with the solute (e.g. via H-bonds) can help or hinder crystallization and may or may not produce crystals containing solvent of crystallization.
3. Solvents can affect the shape of the crystal, though again you can rationalise the effects by saying that interactions are too strong or too weak.
4. Viscous solutions crystallize more slowly.

When decades of work confirm what most of us would have worked out for ourselves, one has to wonder why so much effort has gone into it. To be fair, the bit about viscosity might not be so obvious. Most of the time a difference in a few cP of viscosity makes little difference. When you start cooling solvents that have lots of desirable donor/acceptor properties, they can start to get seriously viscous, maybe slowing down nucleation by a factor of 2. That wouldn't be much if nucleation times shifted from, say, 5s to 10s, but is a big deal if it shifts from 5hrs to 10hrs.

My explanation for the dominance of CNT is "physics envy". I take up this point when we discuss crystal growth.

11.7.4 2-Step CNT

The experiment with "memory" above the saturation limit gives us a feel for what's going on and fits in (though this is seldom mentioned) with KB theory. Solutes don't just go from "happy" above saturation to "unhappy" below. They become increasingly unhappy. If you measured the G_{uu}, the solute's tendency to self-cluster, you would see it steadily increase as you approached saturation. If you have a neutron scattering machine handy[111] you can even go below saturation and see those clusters increase in size.

---

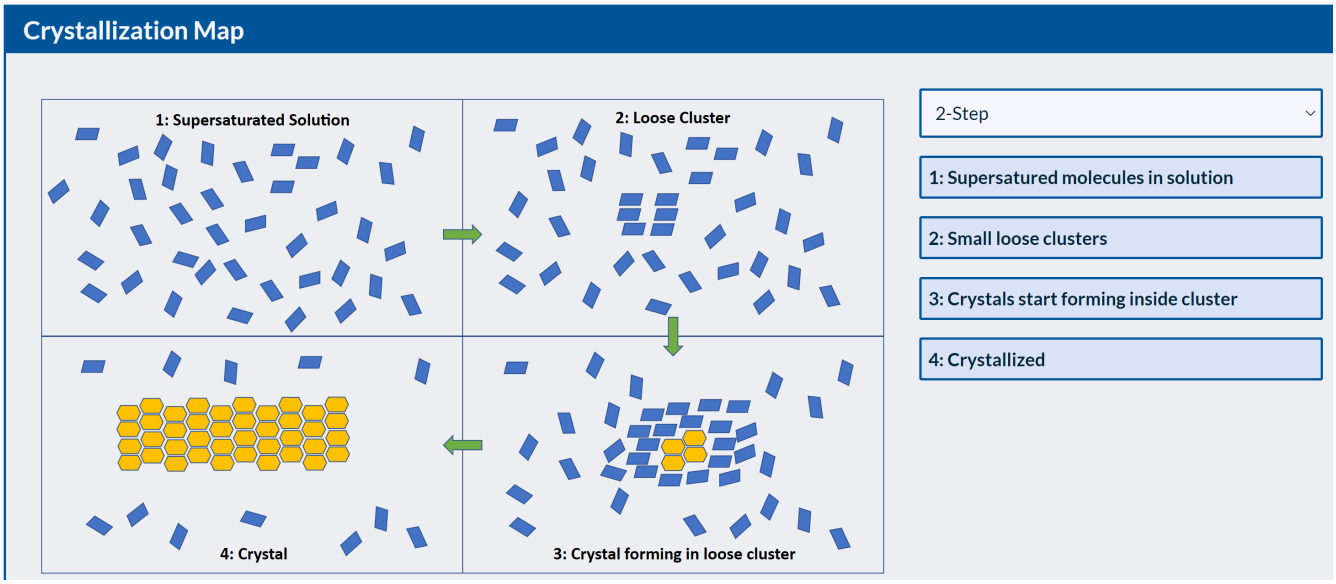
111 Not many of us do and, of course, these experiments can be frustrating if the crystals nucleate before you have done your (expensive) measurement.

This sort of statistical clustering isn't necessarily creating pre-nuclei. As KB theory shows, (and as can be seen in large-scale simulations discussed below) a large G_{uu} can come from lots of very close self-clustering (what we might call a pre-nucleus) or an extended amount of looser clustering. Let's go with the second type and carry on cooling. We now have a very extended amount of looser clustering. If we keep going that loose cluster phase separates into what it was really meant to be - a liquid. Just as immiscible solvents can phase separate when you cool them, so the solute can phase separate as a viscous liquid phase of concentrated solute. This is the frustrating oiling out, discussed in more detail shortly. It's not something pathological; it's one very natural route for a cooling solution to follow.

Of course, if it's the tighter sort of clustering, and if it starts to get nicely ordered, with the degree and direction of ordering depending on solute-solvent interactions, then you can reach a stage where it's more of a crystal than a cluster and crystallization can continue.

If gently heat some crystals, they can dissolve into a reasonably tight cluster just above the saturation limit. Left for a few hours, the cluster will randomize. But left for 30min there's enough self-ordering that on re-cooling the crystal will grow in that same form, with a smaller lag time - that's the memory effect.

It might be that the surface energy cost of the forming crystal being in contact with the solvent is simply too high, while a loose cluster in the solvent is no problem. This opens the opportunity for a crystal to form inside the loose cluster, now with a low surface energy cost because it's surrounded mostly by itself.



App 11-5 https://www.stevenabbott.co.uk/practical-solubility/Crystallization-Map.php

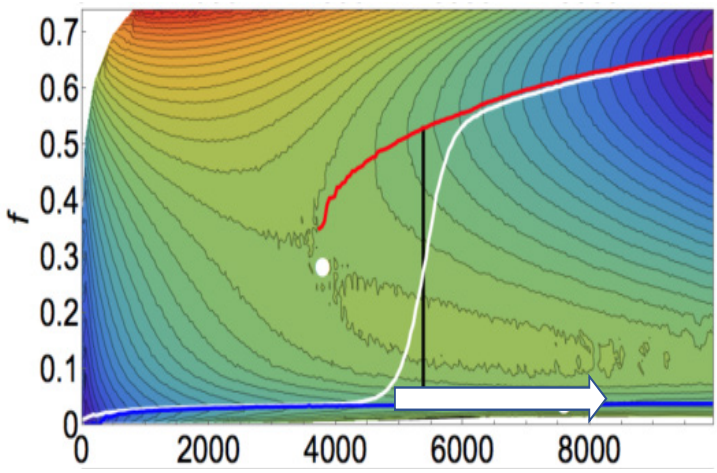
Later we shall properly explore the full range of ideas via the Crystallization Map app. For the moment we can get an idea in the screenshot of the 2-Step mode with internal crystallization.

So 2-step CNT is getting rid of the naive notion of solute molecules doing an all-or-nothing ordered clustering towards a critical nucleus. It's agreeing with a lot of other statistical thermodynamics, making crystallization less of a unique mystery and more of a continuum with other phenomena. It is not a coincidence that 2-step theory has been well-established in the world of protein crystallization. There, it is much easier to use classic light scattering and old-fashioned osmometry and calorimetry to apply statistical thermodynamics to crystallization. If we all had neutron scattering systems in our labs, then classic CNT would have died a long time ago because we would have been able to see KB-like behaviour routinely.

The status of 2-step theory is odd. Some people hate it and go to lengths to show that it doesn't really exist for small-molecule solutes. Others think it is *the* answer to crystallization. My view is that it is a useful sub-set of the broader cluster approach, discussed below, which is the only one which makes sense across the broad range of crystallization phenomena.

11.7.5 Oiling out

Although oiling out is infuriating, there's nothing mysterious about it. In Chapter 4 we had a KB analysis of t-BuOH, that experiences huge clustering effects in water, bordering on the insoluble. For t-BuOH the balance of thermodynamics happens to keep it soluble; the other butanols show similar behaviour, but their clustering gets too much and they phase separate. Whether a cluster crashes out is down to a few kJ/mole, tiny energies when the clusters are in the 100s of kJ/mole. To switch to a crystalline form undoubtedly leads to a happier enthalpic state, at the cost of a huge reduction in entropy. There can be a thermodynamic mountain to climb before switching to the crystalline form, just as the standard CNT shows there's a mountain to climb before a critical nucleus size is reached. And just as a high viscosity from a relatively concentrated, cold solution can slow down standard nucleation, it can slow down the transition of a large cluster to the many smaller ones, of a different configuration, that can form proto-crystals.



Here the white line shows a trajectory of a 2-step crystallization where the x-axis is the size of the loose clusters and the y-axis is the size of clusters in crystallized form. The white arrow shows that the loose clusters might equally have continued growing in the x-axis direction ultimately leading to oiling out. The source of the image is discussed below,

The stuff that oils out isn't pure glassy material which might then be expected to crystallize as it would do from a melt. It brings a significant fraction of the solvent with it, presumably clustering into its own domains that get in the way of proto-crystals forming. The oiling out is the spinodal decomposition discussed above, and the composition of the two phases might be something like 60:40 and 10:90, with the phase ratio governed by the initial concentrations. These specific values are taken from a paper[112] that places oiling out explicitly in the context of the metastable diagram and illustrates other aspects discussed below. Interestingly, the solvent is a water:ethanol mixture and the separated phases also contain different amounts of the two solvents.

The reason for stressing the boring normality of the oiled-out state is to encourage a fresh approach to dealing with it. Hoping, as I often did, that it was a pathological freakish state that could sort itself out doesn't help. Try some hypotheses. If the molecule has two donor sites and the solvent is an acceptor, maybe there is no chance of somewhat ordered pre-crystals to form. By adding a modest amount of donor co-solvent, that might out-compete the donor-acceptor interactions of one site, it might be possible to tilt the subtle balance in your favour. Via COSMO-RS or maybe just HSP you can get a sense of which co-solvent might give you the effect you desire without unduly disturbing the solubility balance you carefully set up in the first place. If you only have a small mole fraction of solute, you don't have to add much of your specific co-solvent to out-compete it. If you have 1g/l of 180 MWt molecule, simple arithmetic shows that you would need to add only 0.1g/l of water to be a molar equivalent.

Which is a good reminder that for molecules that you are trying to control via subtle donor/acceptor balances, you'd better make sure you dry your solvents so that your careful plans aren't being outsmarted by "small" amounts of water.

11.8 Crystallization Map and Nucleation

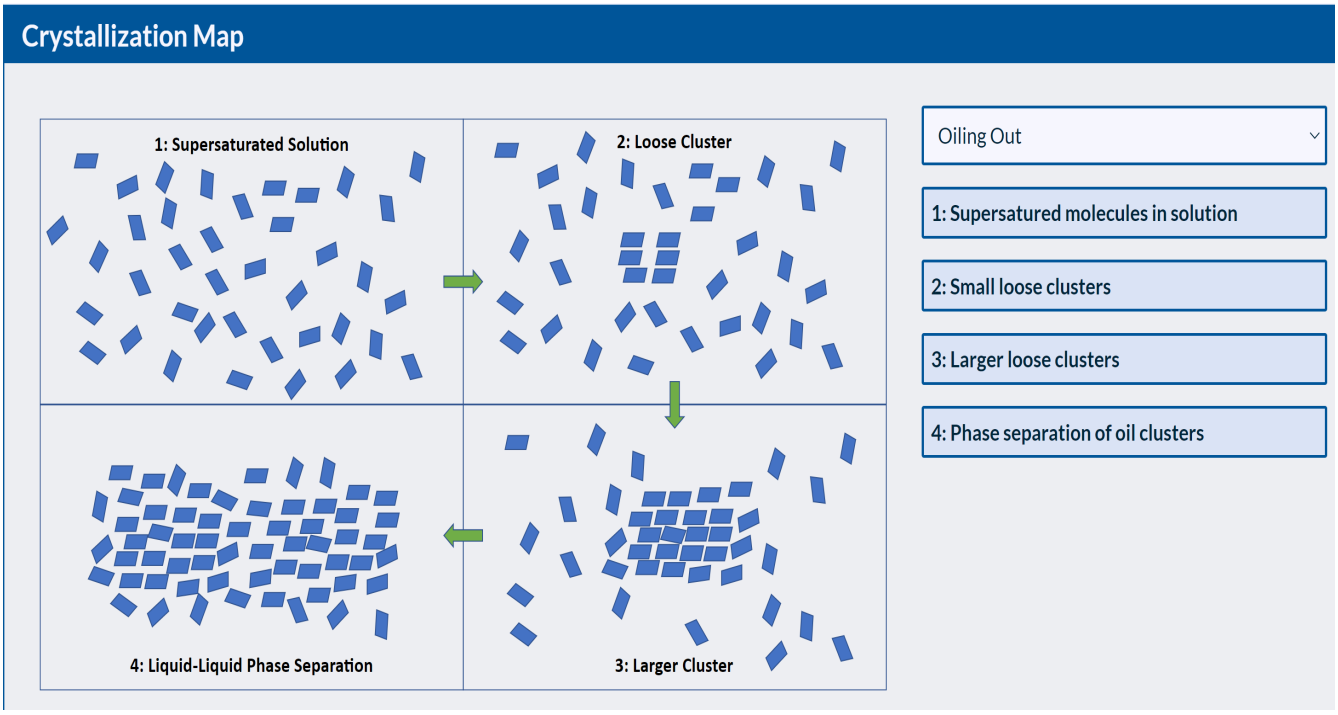
One of the striking features of the crystallization literature is that you can read a perfectly reasonable explanation of how crystallization really works, then go to papers citing that explanation and showing that it's wrong via a few counter-examples. The best explanation for this diversity of views seems to be that there are multiple possible routes from supersaturated solution to a given crystallized form and that the conflicting explanations focus on just one type of route.

To put it another way, crystallization is a complicated landscape of mountains, valleys, saddle, cols, side valleys, with different routes leading to different final destinations only one of which is at sea level, the thermodynamic optimum, with others being stable regions far from the sea or in different watersheds.

---

112 Emilie Deneau and Gerry Steele, *An In-Line Study of Oiling Out and Crystallization*, Organic Process Research & Development 2005, 9, 943-950

If you have a supercomputer and some plausible model of the energy landscape you can produce fascinating maps of how crystallization might depend, for example, on the relative (effective) surface energies of clusters or clusters within clusters. Why "clusters within clusters"? It's possible that one arrangement of the solute molecules is very high surface energy with respect to the solvent, yet it is low surface energy with respect to the cluster form of solute+solvent in some part of the energy landscape. This idea is most associated with ten Wolde and Frenkel[113] in the context of proteins, and has been developed in various directions since then. A paper[114] which covers the modern simulation tradition from Prof Poole's group in St Francis Xavier university does just this and inspired the app that will follow shortly (and the map image earlier, kindly provided by Prof Poole). It is based on fluctuation theory and therefore fits into the KB background to this book. Although the paper contains the outputs from high-powered simulations, at the core is the fact that it is the surface energy of clusters that dominates the pathways, so a favoured final form which has a higher surface energy may not be reached directly - instead it will appear as a sub-cluster within a lower surface energy cluster.



App 11-6 https://www.stevenabbott.co.uk/practical-solubility/Crystallization-Map.php

The app makes no attempt to draw complex 3D energy topographies. Instead you get to choose one of a set of potential pathways to see what's going on. Earlier we saw the 2-Step crystal mode. In the example shown here, the loose

---

113 Pieter Rein ten Wolde and Daan Frenkel, *Enhancement of Protein Crystal Nucleation by Critical Density Fluctuations*, Science, 277, 1997, 1975-1978

114 Daniella James, Seamus Beairsto, Carmen Hartt, Oleksandr Zavalov, Ivan Saika-Voivod, Richard K. Bowles, and Peter H. Poole, *Phase transitions in fluctuations and their role in two-step nucleation,* J. Chem. Phys. 150, 074501 (2019)

cluster in step 3 just gets bigger, eventually transforming to a very large cluster - i.e. an oiled-out drop, trapped in a local minimum and never able to crystallize.

A complementary strand of simulations of early nucleation clusters is the series of papers from Lutsko[115] who gradually clarifies his own ideas and accesses more powerful simulation tools to emphasise that the standard idea of a steady accumulation of molecules from an initially homogeneous concentration distribution makes no sense. It is large-scale density fluctuations that are the key, and if you happen to have (in my language, not his) a G_{uu} reflecting a large loose cluster, this can readily become a similarly large G_{uu} but as a much tighter cluster with a smaller effective radius, before growing again to a cluster that might easily oil out and grow or internally crystallize and grow. The freely downloadable video from of this latter case (video S4) is a joy to watch.

Wonderful though the map is, without some ideas of how nucleation *really* works, it's largely irrelevant because most realistic processes include seeding. To put it another way, we have been thinking about pure homonucleation. As far as I'm concerned, there is no substantial mechanistic difference between heteronucleation (nucleating on junk), secondary nucleation (nucleating on seeds created by growing crystals) and direct seeding (seeding onto externally added crystals)[116]. There are many *system*-wide differences between these as they give various levels of control, but for the moment I want to focus on seeding as a general phenomenon, because a rather straightforward picture emerges, building on the ideas behind the crystallization map.

11.8.1 Seeding

I had always assumed that seeding was straightforward. You had a supersaturated solution that would take forever to nucleate spontaneously, so you dropped in a seed crystal and things looked after themselves. It turns out that seeding is as confusing as everything else about crystallization.

If you read old literature on seeding[117] there are many ingenious experiments which seem to answer some specific questions for a specific solute, then you go onto another paper that addresses a different set of issues. Fast forward to 2021 and the situation basically hasn't changed.

A seed with a surface that has been carefully pre-cleaned may (or may not) behave very differently from a seed that (for whatever reason) contains lots

---

115 A good specific example, with a wonderful video, is James F. Lutsko, *How crystals form: A theory of nucleation pathways*, Sci. Adv. 2019;5: eaav7399

116 The papers of Botsaris, discussed below, are full of important experimental details. One paper notes with great surprise that Teflon shavings, accidentally added due to their admitted poor original experimental design, are quite effective at nucleation - not as good as adding real seed crystals, but not at all bad.

117 Once again I am grateful to Prof Threlfall for his capacious knowledge of the literature back into the 19th century, even in the original German.

of "powder" on the surface. Seeds must be exactly the right crystal form or, alternatively, can be any bit of junk floating in the solution. And, of course, we have those infinitesimal, near-magical seeds that move from lab A that produces polymorph Y to lab B which only produced polymorph X but from the day the visitor from lab A arrived can never again produce X.

Stirring too much or too little is bad. Zapping with ultrasound or lasers may help or hinder.

The obvious notion that you just throw in some seed crystals which then happily grow isn't obvious at all. We have papers assuring us that seed crystals must not be too big nor too small, with zero agreement on what constitutes an optimum size. Generally there follows a discussion saying that via Ostwald-Freundlich theory (we met this in terms of nanoparticles in Ch 7) crystals smaller than, say, 1μm will dissolve because of their high surface energy, but that begs the question of how any ordinary crystallization can happen.

Throw in large seeds and they also do nothing - following the common observation that for reasons unknown (put down to vague "build up of impurities") many crystals reach a natural maximum size. Throw in large seeds so they can break down into smaller ones, then we may get crystal growth, though there's a contradiction with the notion that small crystals are too small to be able to seed.

The question is complicated by the fact that in plenty of crystallizers there are "fines removal" systems that extract useless small crystals and put them back to redissolve in the starting liquor. These are big enough to see and small enough to be a nuisance. Somehow they aren't taking the system along the path to the desired crystal size.

One type of experiment (referenced later) takes a solute such as sodium chlorate that grows chiral crystals. Randomly seed it and you get 50:50 chiral forms. Take one of the forms as the seed and do various things to it, you might get 100% of that form - showing that it must be cloned chunks that have done the seeding. This is the classic "crystallization is exquisitely controlled" theory. Or under different circumstances you might get 50:50 showing that despite the fact that the seeds you added were chiral, they are not exerting any control.

Because for any production crystallization you need good stirring for uniformity of composition and temperature, seeding takes place in well-stirred systems and is generally more successful. Because crystals can bash into each other, the mixer blade and the walls of the container, it's natural to assume that the seeds are broken fragments. This is the classic "attrition" cause of nucleation - i.e. the added seeds are merely sources of smaller seeds smashed off the larger crystals. But there's plenty of evidence that malformed fragments with lots of strain energy aren't great seeds; indeed, that extra surface strain might

be enough to cause them to re-dissolve leading to seeding via local higher supersaturation.

In any case, for the smaller crystals generally preferred by pharma, this attrition simply does not, *cannot*, take place. The reason (a "squeeze flow zone") is discussed below.

Then there are some lovely experiments where seeding happens by gently sliding the crystals along the bottom of the container or by tapping with a glass rod. In both cases there is no obvious damage to the crystal.

Because there are strong traditions of crystallizing generally simple salts and sugars, alongside traditions of crystallizing pharma compounds, mechanisms that might plausibly work for multi-molar concentrations of easily crystallized inorganic solutes seem unlikely to be directly transferable for modest concentrations of slow-seeding, slow-growing organic crystals.

Sometimes you need to add 2% of seed crystals to get a batch to grow properly, sometimes, as we've seen, it needs one visitor from a production site that produces polymorph Y for the new site to shift forever from polymorph X to Y.

As it's always good to attack one's own biases, I set out to disprove the idea that "any bit of junk" can act as a seed. It turns out that there are plenty of papers whose title implies that you really can seed effectively by adding, for example, polystyrene nanospheres. The title and abstract are encouraging. Then you get to the experimental data. For some solutes there is a modest increase in the number of crystals, for others no change and for others a modest decrease. In the end each paper concludes that too much of anything is bad - a fact both true and unhelpful. Those who like a little more precision say, for example, that a strongly hydrogen bonding surface might encourage hydrogen bonding solutes to more easily form a big nucleus and grow to a crystal or, exactly the opposite, might attract the solute sufficiently for it to prefer the surface to more solute, thereby inhibiting crystallization. So again they are saying that too much of a good thing is a bad thing, true but unhelpful.

Another favourite theme is to change the "surface roughness" or "surface structure". And the results continue to be ambiguous - some changes to the surface, or some specific surface shapes (squares, triangles, wedges...) are shown to help or hinder in almost equal proportions.

Both the "different surface" and "different shape/size" types of experiment lack two crucial ingredients. The first is a lack of the sort of scientific hypothesis that can be tested meaningfully. Most of them say that there is a critical nucleus size and that their specific surface may, or may not (invoking "hydrophobic" or "hydrogen-bonding" effects), make it easier for that critical size to be reached, arguing in both directions. Some of the surface shape hypotheses are slightly

more refined saying, for example, that a specific crystal form would have difficulty seeding on a specific surface shape, such as a wedge.

The second is a lack (the opposite problem is discussed shortly) of a sense of appropriate scale for the experiments. If, as is commonly assumed, a critical nucleus size is 50 or 100 molecules, then the relevant sizes are in the few nm domain so surfaces/shapes need to be characterised and compared/contrasted at the sub 10nm scale, a task so difficult that few papers even discuss the possibility. A surface containing a few wedge shapes is *not* in general going to contain sufficient wedges of sufficient acuity to be able to have a statistically significant effect in a system that is likely to contain 10^{16} nano-sized bits of junk floating around. Surfaces with varying roughness are popular because our intuitions (from hand-drawn diagrams) is that it is trivial to get a massive increase in surface area with a bit of sandpaper. A bit of thought aided by an app (https://www.stevenabbott.co.uk/practical-coatings/surface-profile-explorer.php) shows that most of the time we add, at most, a few % surface area by roughening. However, that sort of argument applies only to macro-roughness in the, say, 100+nm domain. What about roughness at the nm scale relevant to seeing? The papers I've read don't even raise the question; a "rough" surface is good enough for them.

There is a wonderful research tradition where these nano-scale challenges are fully embraced. Via a variety of nano techniques it is possible to make exquisite structures with controlled shape (round, square, hexagonal ...) and size (5-100nm) and then to watch crystals grow inside, sometimes faster than on a flat surface, sometimes slower, sometimes with weird polymorphs forced on the crystal by the confinement. Now the problem is that it is hard to see how to apply the lessons from these experiments to our own crystallization issues. The experiments confirm a lot of what we would expect without allowing us to implement the insights in our macroscale tubes or vats.

What this all means is that you can find definitive experiments that prove just about any mechanism and disprove others, and then equally definitive experiments that prove the opposite. And that's just with seeding, what about crystal growth?

11.8.2 Crystal growth

Not surprisingly, given that we can't agree on how seeds start growing, we can't agree how crystals grow. Standard nucleation theories say "Phew, we've gone past a critical size now the surface energy effects are overwhelmed by the bulk crystallization effects, so crystal growth just happens". Yet crystals can be frustratingly slow to grow. There are many schemes offered to explain why the solute molecules, known to be thermodynamically happy once joined to the crystal, join so slowly. I think the evidence is strong that most of these schemes are wrong. That's not such a shocking statement because there is

general agreement that most schemes historically proposed for crystal growth are, indeed, wrong. The BCF (Burton, Cabrera and Frank) is the most popular explanation I've found, even though I can find no compelling arguments that it is of any use to our community. The idea is a bit like CNT - the molecules can only assemble if there's something to assemble on. A smooth crystal surface provides no high-energy site for a molecule to lock on to. "Therefore" the surface must be rough. Obviously, if molecules lock on to the rough portions, there's a danger that the surface will become smooth. So BCF say that there are screw dislocations which just carry on growing forever. Others say that actually surfaces can be rough for various hand-waving reasons so we don't need BCF, but can use whatever numerical tricks BCF offers to understand crystal growth.

Earlier I mentioned that CNT is a sort of physics envy idea. The same applies to BCF. CNT developed out of nucleation of things like rain drops. The theory makes sense here because a water molecule in the vapour phase and one in a (proto)drop are very different. Similarly, BCF can be observed on atomic crystals growing via vapour deposition. There's nothing wrong with CNT and BCF in those circumstances. But a supersaturated solute in a solvent has a much larger set of choices than an isolated water molecule in the sky or a gallium atom nearing a cooled gallium surface. As we have seen, in our world there isn't a yes/no choice between "seed" and "nothing", there is a vast array of possibilities. Similarly, as we shall see we can have clusters sitting on the surface of a crystal with the choice of being, or not being, incorporated into the crystal itself. This is a very different universe from the imagined ones of single molecules trying to attach to the surface. The physics envy casts the discussions in the wrong framework. No wonder we have problems trying to make sense of them.

In fairness, the standard reference work on nucleation and growth from Kaschiev[118] includes the more relevant "polynuclear polylayer" idea where what I would call a cluster in contact with the surface undergoes a combination of spreading and heteronucleation (yes, that's the term he uses), resulting in the newly-crystallised molecules becoming part of the larger crystal - with a shape and size that depends on the initial cluster size and the relative speeds of spreading and crystallizing. The formula (his Equation 27.21) for the crystal growth rate contains only 4 terms which need to be expanded and rapidly spin out of control as you head for (27.31) which contains 8 terms, one of which is itself made from 7 terms. As most of us can't access 14 parameters to model polynuclear polylayer growth maybe it's not surprising that this theory seems not be be well-known. Because "heteronucleation" is conventionally meant to be about the starts of seeds around bits of junk, it is surprising (though logical)

---

118 Dimo Kaschiev, *Nucleation, 2000, Elsevier.* I was dismayed to find that Kaschiev defines clusters as having a clear boundary between inside and outside - exactly what I do *not* intend. He contrasts his "clusters" with "density-functionals" i.e. clusters with no definitive boundary. As he pleasingly points out, density-functionals have the advantage of being fundamentally right while his cluster theory is objectively wrong. Their disadvantage is that performing meaningful calculations with them is currently not viable.

to see it meaning the merging of a cluster with an existing macrocrystal. It is another reason for abandoning the distinction between nucleation and growth.

The net result is that after decades of theorising we are left with crystal growth formulae which tell us that the growth rate G depends on the current crystal size L, supersaturation S to some power x, and a couple of constants, one of which, k_r describes relative diffusion and incorporation rates:

$$G = \frac{kS^x}{1 + k_r L}$$

Equ. 11-3

Although this is fine as a general-purpose phenomenological equation, it's just a handy formula that any of us could have invented. It has proved impossible to work out in any meaningful way how to use k, x and k_r to gain deep insights into what is going on.

There is *some* logic to k_r as it is trying to capture the relative ease of the solute diffusing to the surface compared to being incorporated once it is there. Here again, the wrong image of what is going on has caused more confusion than insight.

The core problem is that there is a contradiction at the heart of the "diffusion versus incorporation" issue. There is no doubt that, for unstirred situations, there is a concentration gradient from bulk to near crystal. There are beautiful interferometric measurements of these gradients[119] with fascinating effects "at" the crystal surface, but these are misleading because the experiments are looking at effects across distances in the multi μm range and surface effects are in the nm range[120]. There are also plenty of discussions about "boundary layer" effects that drive experts mad[121] because the phrase is used too loosely and causes great confusion. However, there is also plenty of evidence (and Kaschiev takes it for granted) that clusters of solute like to sit on the surface of crystals, so we have the odd situation of an excess concentration (the clusters) at the location where gradients and boundary layers say there is lower value, equal to the saturated concentration.

It is these clusters that are the key to restating what's going on in crystallization. In many, perhaps most, reviews of crystallization, they are simply not mentioned. Yet they aren't a secret. The Botsaris ECSN (Embryos Coagulation Secondary Nucleation) idea has been around in tentative form long before its 1997

---

119 The 1949 paper by SPF Humphrey-Owens, *Crystal growth from solution*, Proc. Roy. Soc. A , 197, 218-237, 1949, is a delight to read and a reminder that sophisticated experiments aren't only do-able in the 21st century.

120 Many problems in science arise because our sketches of what's going on are hopelessly out of scale. The image takes our minds off in scientifically implausible directions, and we are unaware that this is happening.

121 The title of Franz Rosenberger, *Boundary layers in crystal growth: Fact and fiction*, Prog. Crystal Growth & Character. 1993, 26, 87-98, shows this frustration.

description[122] and 1998 & 2004 illustration[123] using chiral seeds in sodium chlorate, as mentioned above. The idea is that CNT-style clusters can't naturally grow big enough to make good seeds, but these "embryos" (hence the name) or "colloids" can be attracted by vdW forces to a large seed crystal where they can assemble into larger colloids that, in turn, can either become part of the crystal or be swept off the crystal as fresh seeds. At lower saturations, the fresh seeds seem to be strongly influenced by the big seed so the resulting crystals are the same chirality as the large seed. At high saturations, the large seed is provably generating many more fresh seeds, but they must be clusters being attracted and assembled from the bulk, with little influence by the surface, before being swept off, as the resulting chirality is random.

Although this section is about crystal growth, we have to make a diversion into this topic of secondary nucleation. We currently don't have techniques for telling us what fraction of these clusters end up as part of the crystal and the fraction that gets swept off to become seeds for fresh crystals. This will be different for each solute/solvent pair and depends strongly on other aspects of the system[124]. It is obvious that the proportion swept off will depend (if researchers aren't carefully sliding the crystals in demonstration experiments) on the local turbulence - a fact confirmed with experiments discussed shortly. The other factor is sometimes called the SNT, Secondary Nucleation Threshold, shown in the diagram near the start of this chapter. At concentrations below it, there aren't enough clusters on the current crystals to be swept off, so the current crop of crystals keep getting larger. Too far above it and other effects such as homogeneous nucleation or oiling out kick in. For pharma crystals in a cooling batch crystallizer there is, inevitably, a balance between keeping the concentration above the SNT and continuing to produce (mostly desirable) small crystals, and making sure that there aren't too many fine crystals that could mess up downstream processing.

A nice simulation of the *principle* of clusters assembling on a surface comes from Anwar's group with the ambitious title: *Secondary Crystal Nucleation: Nuclei Breeding Factory Uncovered*[125]. As with many other simulations, Lennard-Jonesium is used, with the usual fine-tuning of relative like/dislike of solute and solvent particles. With appropriate conditions they could get conventional

---

122 Ru-Ying Qian and Gregory D. Botsaris, *A new mechanism for nuclei formation in suspension crystallizers: the role of interparticle forces*, Chem. Eng. Sci., 52, 3429-3440, 1997

123 Ru-Ying Qian and Gregory D. Botsaris, *Nuclei breeding from a chiral crystal seed of $NaClO_3$*, Chemical Engineering Science, 53, 1745-1756, 1998; Ru-Ying Qian, Gregory D. Botsaris, *The effect of seed preparation on the chirality of the secondary nuclei*, Chem. Eng. Sci., 59, 2841-2852, 2004

124 Botsaris shows that unwashed seeds spray random dust, that seeds washed in unsaturated solution have a "dissolving crystal" surface which behaves differently from those washed in a saturated solution. Seeds at a lower temperature cause local high supersaturation and those at higher temperature again give a "dissolved" surface.

125 Jamshed Anwar, Shahzeb Khan, and Lennart Lindfors, *Secondary Crystal Nucleation: Nuclei Breeding Factory Uncovered,* Angew. Chem. Int. Ed. 2015, 54, 14681 –14684

molecule-by-molecule crystallization onto a crystal already present, then, with some tweaks, they could get clusters forming both on and around the surface, the "breeding factory". These clusters in turn could be imagined (MD simulations can't do turbulent flow) to be easy to sweep off (secondary nucleation) or incorporated into the main crystal (crystallization via cluster contact). Although Lennard-Jonesium is simplistic, it shows the core principles irrespective of specific molecular interactions. The paper was grounded in the real world because there is a beautiful SEM of a real pharma API where the surface clusters on an original seed had stayed in place and crystallized. The authors also note the idea from Van Driessche of "closed loop macrosteps" which can conveniently be translated into "blobs sitting on the surface". We meet these looped macrosteps later as they can sweep away impurities at the crystal surface.

A paper[126] that explicitly brings Anwar's ideas to life is from de Souza and colleagues. It carefully looks at secondary nucleation of much-studied paracetamol and beautifully demonstrates that all the secondary nuclei must be produced via clusters on the surfaces of the larger seeds added as per conventional batch nucleation. They can eliminate the common idea of seeds being produced via crystals bashing into things. They use indirect means (no evidence of smashed crystals) and direct experiments that attempt to smash crystals into a solid surface only to find that they get deflected away by a "squeeze film" of liquid between the crystal and the wall. A follow-up paper explores the science of squeeze films further.

Why does de Souza find no attrition and yet many others find it easily? Pharma crystals tend to be in the few 100µm range (for easy conversion into tablets and for relatively rapid dissolution) while lots of classic crystal studies are in the mm range. The physics of smashing into impeller blades is very different for those large crystals. Crystallization is very much a property of the system and you cannot just take an idea from one system (smashing large crystals) and apply it to another (small crystals incapable of impacting a surface at any reasonable speed).

It is also taken for granted in some reviews that it is loose clusters (whether assembled via vdW forces or not, not everyone knows of ECSN) that are swept off a seed crystal surface by sliding the crystal or merely tapping with a rod, neither of which causes classic attrition of the crystal surface. However, because this production of seeds is often called "contact nucleation" there is a bias towards the "contact", which (think of the brutality of "contact sports") then blends into attrition ideas, rather than towards the fact that the contact takes away loose clusters from the surface. If contact nucleation had been termed "cluster nucleation" from the start, a lot of confusion could have been

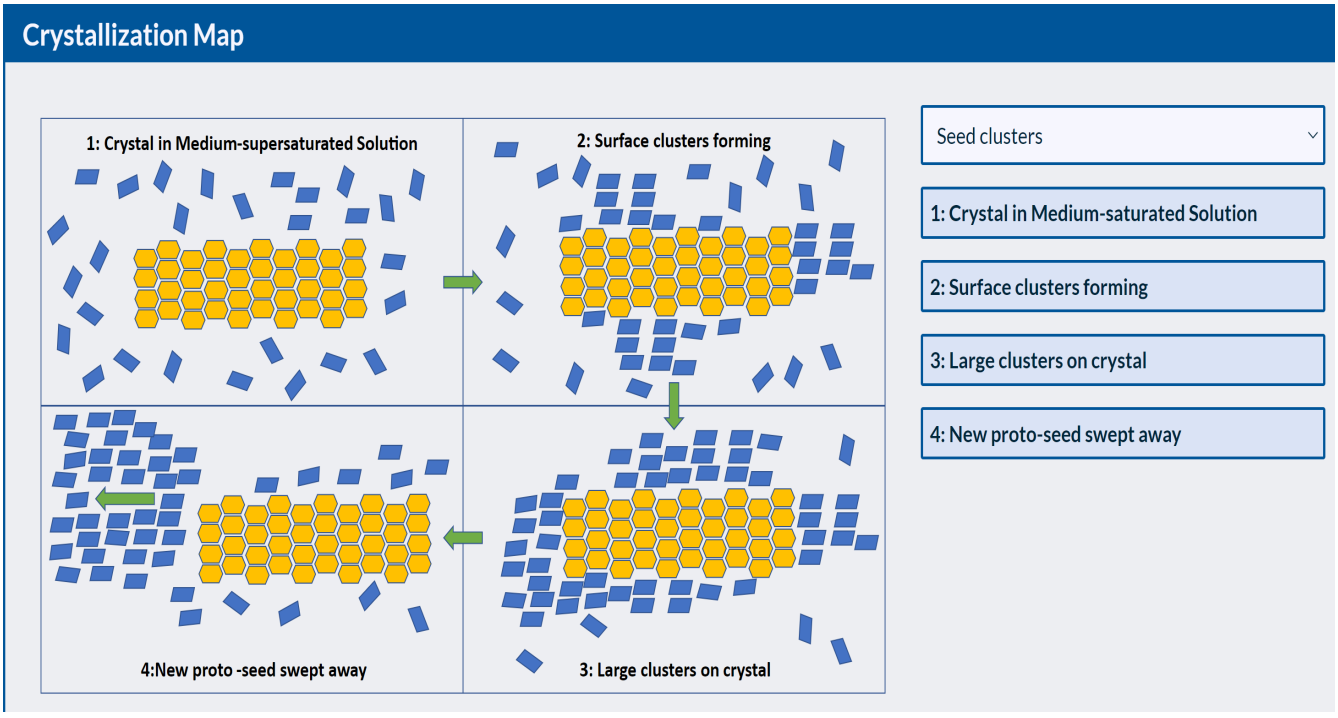
---

126 Brian de Souza, Giuseppe Cogoni, Rory Tyrrell, and Patrick J. Frawley, *Evidence of Crystal Nuclei Breeding in Laboratory Scale Seeded Batch Isothermal Crystallization Experiments*, Cryst. Growth Des. 2016, 16, 3443−3453. Later papers from the same group refine the story but do not change it substantially.

averted. My one criticism of the de Souza paper is that they choose to call the mechanism "crystal nuclei breeding". By using yet another term for the same phenomenon (and one that could still apply to attrition), the momentum behind cluster nucleation is reduced. Others from the same research institute say that "crystallites" are swept off the crystals. To me a "crystallite" implies something that is already crystallized, so the term is pre-supposing a mechanism. I keep preferring the neutral term "cluster" until we have definitive data to qualify it. Maybe some secondary nuclei are "liquid clusters", others are "amorphous solid clusters" and others are "semi-crystalline clusters" and others are "perfectly crystalline clusters". I would be surprised if we could not find examples of each. Till then, let's not close off discussion by using a name, "crystallite" that presumes (at least in my interpretation) something that is far along the continuum from liquid to crystalline.

Still, these reviews are more interested in the secondary nucleation effects than in the clusters, so even when clusters are assumed, they are merely part of secondary nucleation. The obvious question "If there are clusters sitting on the surface, why aren't they part of the crystal" seldom gets asked. Or even if it does get asked, the answer, as per Kaschiev above involves 14 parameters.

The Crystallization map app tries to tame this confusing whirlwind of options via a set of specific examples. Let's look at core secondary nucleation:



App 11-7 https://www.stevenabbott.co.uk/practical-solubility/Crystallization-Map.php

The image shows the surface clusters growing (by whatever mechanism, maybe vdW attraction of smaller ones) then a proto-seed being swept away hopefully to start its transformation into a new crystal.

Other options you can select from the app are classic crystallization where individual molecules just arrive and get incorporated, and the other extreme where clusters grow too quickly in a disorganized manner to give dendrites.

There's another interesting option. Suppose the clusters are nano-sized. That's large by the standards of CNT but small compared to a real crystal. What would happen if those nano-clusters just crystallized out, one after the other, with adequate mutual alignment (so in x-ray they look like a single crystal) but with clear boundaries between them? We know the answer because Van Driessche[127] has some beautiful images of calcium sulfate crystals formed in exactly that manner. This isn't something you can see casually - it's a lot of work with powerful electron microscopes. I expect that it's quite common, with people not looking for it because standard molecule-by-molecule assumptions make such a crystal unimaginable. Loosen the constraints a little so that the nanocrystals are allowed to be more easily visible and we have the more easily identified "mesocrystals" common in nature because the gaps can be empty or filled, and allow a wide variety of functionality beyond just being a lump of crystal. The app uses one Mesocrystal image to describe both types.

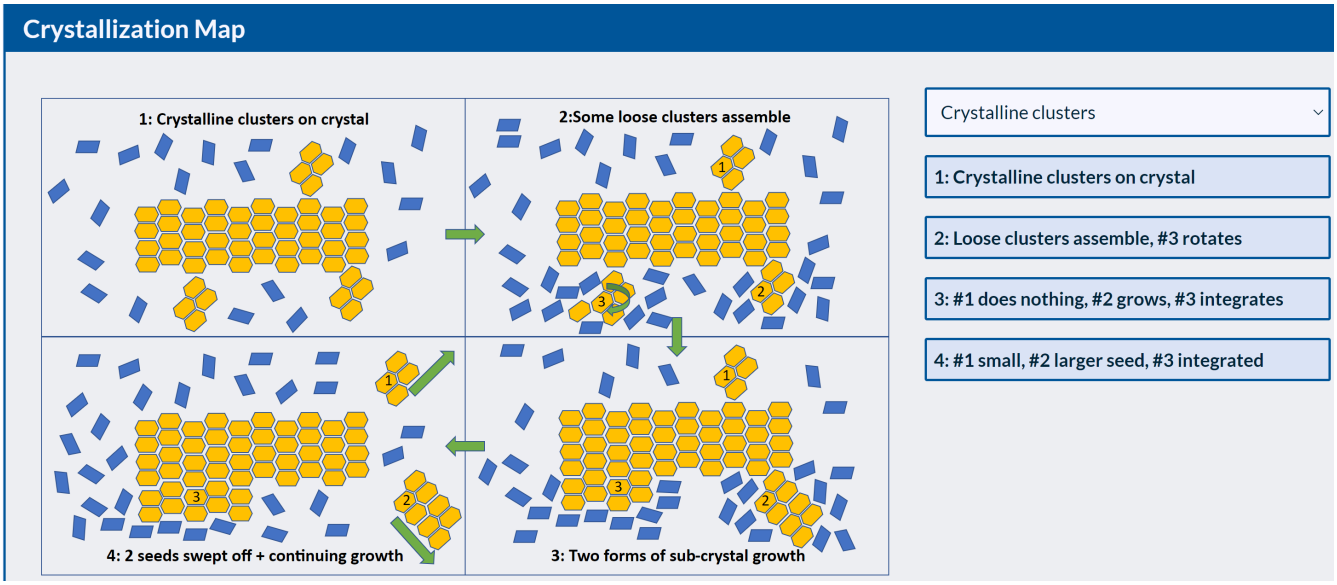
There's a bonus idea that emerges from these clusters sitting on the surface. Suppose there's an impurity at the surface, inhibiting growth. If this impurity area gets covered by a cluster and the impurity now finds itself in an environment where it is more soluble, it can detach from the surface while the cluster itself integrates onto the surface, kicking out the impurity. In the cluster language, again, of Van Driessche[128] but, interestingly, in the different world of protein crystallization, this is "looped macrostep mediated self-purification". The phrase looped macrostep means nothing to most of us, but the paper kindly explains that in real language it is called a 3D island, and, as mentioned above, in my language they are simply a blob sitting on the surface, as simulated by Anwar, and linked to their long-term fate of being converted (in that specific case) into mini-crystals sitting on the original seed. Note that in other circumstances, the looped macrosteps can bury the impurity, allowing crystallization to continue.

That raises one extra case. Suppose the clusters crystallize before they land on the surface. What happens? It's not obvious how they could integrate into the crystal. So do they just wait to be swept off, or do they pick up clusters that have been attracted by the large crystal (if, indeed, that happens) and if so, do the clusters allow the small and large crystals to become integrated? There are so many logical possibilities, as the Crystalline clusters option tries to show.

---

127 Tomasz M. Stawski et al, *Particle-Mediated Nucleation Pathways Are Imprinted in the Internal Structure of Calcium Sulfate Single Crystals*, Cryst. Growth Des. 2019, 19, 3714−3721

128 Mike Sleutela, and Alexander E. S. Van Driessche, *Role of clusters in nonclassical nucleation and growth of protein crystals*, www.pnas.org/cgi/doi/10.1073/pnas.1309320111



App 11-8 https://www.stevenabbott.co.uk/practical-solubility/Crystallization-Map.php

As in so much else, Botsaris is well aware of the question of whether his embryos are liquid-like or crystalline and concludes, wisely, that it depends on the specifics and that he's in no position to be able to conclude one way or the other, though in general for his system the liquid-like clusters are the more likely.

As was mentioned earlier, if we see "turbidity" the assumption seems to be that these are crystalline scattering units rather than fluid clusters. It seems to me to be likely that there will be many examples where the turbidity proves to be from fluid, rather than crystalline clusters. This doesn't have to be "oiling out". Readers who recall the pre-ouzo clusters from Chapter 6 will know that these can create visible scattering but are still in a single phase.

11.8.3 Growth versus dissolution

There is an interesting discussion in Mullin's Crystallization book[129] about the difference, in well-controlled experiments, between growth rate and dissolution rate. The classic theory implies that the rates should be similar, yet dissolution is often 3-5x faster. Mullin dismisses some standard non-explanations but does not offer convincing alternatives. Yet if you dig into Mullin's own papers where he investigated the effects, you find reference to an idea from Gilmer and Bennema in 1972 based on some crude simulations and called "nucleus on nucleus" growth. In my language that would be clusters on clusters. I'm biased, of course, but clearly the kinetics of assemblages of clusters to produce growth are going to be different from dissolution by molecules escaping from the surface. In looking for further examples, it was interesting to see that for NaCl, which I believe may well crystallize via CNT, the growth and dissolution rates were almost identical.

---

129 See Chapter 6.3 of JW Mullin, *Crystallization*, 4th Edition 2001, Butterworth-Heinemann

At the time of writing I have nothing more definitive to say. It seems a fascinating avenue for further exploration.

To bring this section to a close, I note that although the title of the Van Driessche paper above includes the word "nonclassical", in the summary chapter of the excellent New Perspectives book[130] which he co-edited, there is a caution that it is unhelpful to distinguish classical from nonclassical. Far better just to talk about correct theories, based on sound ideas linked to good data, with predictive power.

My view is that the correct theory is simply "crystallization via clusters" which covers both nucleation and growth for the good reason that the two aspects are, in general, inter-twined.

11.9 Crystallization via Clusters

Once we start to use the language of clusters throughout the crystallization process, a lot of confusion disappears.

So, first, what is a cluster? It is a deliberately vague word[131] that works across the extremes we find within crystallization. At one extreme we find pure crystalline clusters as loved by CNT. I'm prepared to believe that NaCl crystallizes via the classic CNT method. The rapid growth of the simple crystal structure seems very straightforward - a few ions come together into a very small NaCl crystal, a few more add to it and at a certain point the surface energy penalty is beaten by the bulk effect and a large crystal is formed very rapidly. At the other extreme we have oiling out where the cluster is *not* just a glassy form of the solute which *should* be crystalline. From the basics of binodal/spinodal phase separation, as seen in the Metastable app, we know that it is a phase containing significant amounts of solvent, i.e. where the solute and solvent are interacting with each other so that the solute gets no chance to segregate into regions sufficiently dense to be able to phase change into crystals.

In between are looser and tighter clusters, smaller and larger. Statistically some clusters might manage to get large enough on their own to be able to phase change, a rare homogeneous nucleation event. Or some of these clusters might be quite large, because they are loose, then have the ability for more highly ordered sub-clusters to form inside the loose cluster and to crystallise - this is

---

130 Alexander E.S. Van Driessche, Matthias Kellermeier, Liane G. Benning, Denis Gebauer (Editors), *New Perspectives on Mineral Nucleation and Growth*, Springer 2017

131 As mentioned above, in some traditions, cluster means something with a fixed, known boundary, with my sort of definition being, in their tradition, the more powerful "density-functionals". These have an appealing link to Kirkwood-Buff ideas but are, alas, not implementable in any scheme known to myself.

classic 2-step crystallization. A variant in between this and oiling out is growth of phase-separated drops which then manage to crystallize.

Maybe, as ECSN suggests, loose clusters tend to remain small unless attracted to any large (>100nm) surface by vdW forces that increase local cluster concentrations, allowing clusters to grow to a larger size and either be swept off to create their own crystals or to be incorporated into the larger crystal.

By emphasising the generic vdW idea, we see that heterogeneous nucleation is fundamentally no different from seeded or secondary nucleation. Most texts on heterogeneous nucleation show familiar diagrams of contact angles and say that if they are 90° then things are incompatible, and if they are less then the surface is "wetted". These ideas make no sense within a liquid environment where, by definition, every surface is "contaminated" by solvent, so the contact angles of solvent-solute clusters are going to be in the range of a few degrees, not the 45° to 90° invoked to "explain" different types of heterogeneous nucleation. The surprise from Botsaris that his "inert" Teflon was quite a good nucleator is more to do with the image of drops of water sitting on Teflon in the air than the modest competition between the clusters and the solvent for a surface that neither much likes.

A very good question about a random junk surface is "Why should the cluster be inspired by the surface to change into crystal form?". But given that most growing crystals tend to have uncrystallized clusters sitting on *their* surfaces, we have the opposite question "Why do we *ever* have clusters on crystal surfaces when they should just become part of the crystal?" Both are good questions that are left unanswered because few in the community have known to ask them.

In a few paragraphs, the whole of the crystallization landscape is described in a single language of clusters. The thermodynamics of clusters is naturally described in terms of Kirkwood-Buff fluctuation theory and in principle at least is amenable to rather simple measures of KBI derived from densities and osmolalities and vapour pressures. Unfortunately in exactly the zone of densities and osmolalities we are interested in, the changes are rather small so are hard to measure with sufficient accuracy. Because fluctuations are naturally captured by neutron, x-ray or light scattering it is possible in principle to flip between the scattering and the KBI viewpoints. As mentioned before, the "difficult" problem of protein crystallization has been well analyzed because the classic KBI and scattering analyses work well with such large molecules. It requires more effort to gain the same insights for small molecules.

If we can't get adequate data via classic KBI or scattering data, what can we do to better understand what is going on? A beautiful example of an extreme

technique to look at what is going on is the work from the Ito group[132] where a molecule in solution shows one colour of fluorescence, the crystal shows another and the amorphous phase shows a third colour. The video you can download from the journal site (paper and video are open access) is fascinating - you see the amorphous phase appear rather suddenly (presumably a spinodal phase change containing relatively little solvent) then the crystal itself seems to burst out of the amorphous phase. Note that we are looking at 5mm drops so this has nothing directly to say about small clusters - it's just a fascinating reminder that crystallization spans a vast array of possibilities.

For more conventional systems, more imagination is required. As argued elsewhere in this book, using FT-IR or Raman to tease out specific solute-solute, solvent-solvent and solvent-solute interactions can tell you a lot. In addition to the normal use of NMR chemical shifts, relaxation times and inter-molecular interactions via smart pulse trains (e.g. NOESY), there are many possibilities arising from the CLASSIC (Combined Liquid- And Solid-State In-situ Crystallization) NMR technique[133] from the Harris group at Cardiff U, which shows many things going on inside the NMR tube (admittedly one spinning at very high rates which raises a number of issues) in, as the acronym suggests, both the liquid and crystalline states, with the signal from small(ish) seeds being detectable.

Why not look for clusters directly via AFM? Another paper from Van Driessche[134] compares different techniques for measuring crystal growth rates. The AFM technique consistently measured rates 5x faster than the others. The most plausible explanation for this is that the AFM tip is providing local mixing to keep the solute concentration higher at the surface. The tip could also (when they tried deliberately) deform the "soft" surface (their word) and the deformations themselves produced fresh crystal growth. We still have a problem; although AFM is a high magnification technique compared to the others, it is still relatively coarse (at least for typical instruments) when looking for clusters, especially near-liquid ones.

Liquid-phase TEM (LP-TEM) seems the most exciting idea but there are many potential artefacts from the technique (the electrons can do many things such as change the pH of an aqueous system) so it requires a lot of care to be sure the results are valid. The review from Tang's group[135] gives lots of good ideas. My favourite example is the growth of Pd nanocrystals where the TM electrons are

---

132 Fuyuki Ito et al, *Direct Visualization of the Two-step Nucleation Model by Fluorescence Color Changes during Evaporative Crystallization from Solution*, Scientific Reports | 6:22918 | DOI: 10.1038/srep22918

133 Kenneth D. M. Harris et al, *'NMR Crystallization': in-situ NMR techniques for time-resolved monitoring of crystallization processes,* Acta Cryst. (2017). C73, 137–148

134 Alexander E. S. Van Driessche et al, Comparison of Different Experimental Techniques for the Measurement of Crystal Growth Kinetics, Crystal Growth & Design, 8, 2008, 4316-4323

135 Biao Jin, Zhaoming Liu and Ruikang Tang, *Recent experimental explorations of non-classical nucleation*, CrystEngComm, 22, 4057–4073

used to create the Pd atoms from Pd salts. They form loose clusters and then start to crystallize - badly. These bad crystals "breath out" and expel a ring of Pd atoms, then reorganize themselves and breath in. After a few cycles, the crystals are properly crystalline. I can think of no better example of how naive CNT is by comparison. The same review shows the classic case of relatively large (1μm) amorphous calcium carbonate "clusters" spontaneously crystallizing to either aragonite or vaterite. Yet just tweak the system somewhat and you can watch classical CNT happening - vaterite going from 0 to 20nm in 20s.

An exciting hunt for clusters came from a PhD thesis at U Konstanz. The background is interesting. A much-cited paper on PNCs, pre-nucleation clusters, comes from a team including Profs Gebauer and Cölfen at Konstanz[136]. It is much-cited because it is one of those core transition texts which shows the limitations of the past (e.g. CNT and 2-step CNT), and indicates the likely future but doesn't quite settle the whole story. The thesis[137] takes the idea of clusters seriously and looks for them in the crystallization of ibuprofen and of diclofenac by adding HCl to their sodium salts. Just about every possible rational technique is used to detect clusters, along with detailed knowledge of pH, turbidity, isothermal titration calorimetry. Various forms of NMR, electrospray mass spec, Taylor dispersion, dynamical probes, AUC (analytical ultracentrifugation) and cryo-TEM are all used intelligently and self-critically. Sadly, the evidence is mostly elusive, with many promising experiments ruined because crystallization happens on, say, the walls of the ultracentrifuge cell. Some nice indirect evidence came via NOESY NMR on the dense liquid phase that forms from the supersaturated solution. Happily the final attempt using cryo-TEM with diclofenac yields images that I find convincing of the borderline world between clusters, seeds and even bulk crystals with clusters on the surface.

Why don't I endorse the pre-nucleation cluster idea if I'm a fan of the Konstanz work? The name suggests that once nucleation has taken place then everything's fine because we just have crystallization. The evidence is overwhelming that there isn't a distinct line between "nucleation" and "crystallization". If, as I think is certain, there are often build-ups of clusters on the surface of a growing crystal, how they behave is still dominated by cluster physics, though this time with interactions partly into normal solvent space and partly onto crystal surface space. With secondary nucleation, often the clusters are swept off back into the solvent domain. Yet if a cluster gets too big on a crystal it can create its own sub-crystal sitting on the main crystal, as an unwelcome lump. In between, as we have seen, nano-sized clusters can be incorporated into the overall crystal. This all makes sense if we keep to the notation of clusters being the dominant control element. It makes no sense if

---

136 Denis Gebauer, Matthias Kellermeier, Julian D. Gale, Lennart Bergström and Helmut Cölfen, *Pre-nucleation clusters as solute precursors in crystallisation*, Chem. Soc. Rev., 2014, 43, 2348-2371

137 Eduard Wiedenbeck, *Nucleation Precursors of Poorly Water-Soluble Pharmaceutical Compounds*, U Konstanz 2019. Note that my original text failed to point out that NOESY was done on the dense liquid phase.

we try to invoke the idea of pre-nucleation clusters magically becoming post-nucleation entities.

A similar hunt for clusters[138] was more fortunate. This time the solute was OTBN, a cyanobiphenyl, chosen because it is simple, stable and comes (as of the time of writing) in only one polymorph. Put it into chloroform (a good solvent) or methanol (less good) and use DLS, NTA (Nanoparticle Tracking Analysis) or SAXS, and you find, above a certain concentration, clusters in the 10 nm (chloroform) or 50 nm (methanol) size range, more-or-less independent of concentration above a critical value. The chloroform clusters have a central core of chloroform with a ring of OTBN, again emphasising that "clusters" covers a wide range of phenomena. NMR experiments (DOSY) produced diffusion coefficients that went from "large" (i.e. small molecules) to "small" (i.e. large clusters) as the concentration increased. My favourite part of the paper is the effect of agitation of the solutions - the cluster sizes jumped to 100's of nm, presumably because agitation enables small clusters to bash into each other. You would expect that following agitation and the creation of larger clusters, the time for a visible crystal to appear would decrease. This is what they found, with an effect so clear that they didn't need fancy statistics to demonstrate it.

NTA seems especially appropriate for this type of research. For conventional particle sizing its limitation is that it can only handle 10^6 to 10^9 particles per ml, which means lots of dilution. For cluster studies this range is ideal. If we imagine a 0.167 M solution then it contains 10^{20} molecules/ml. If we had 10^8 clusters (the sorts of numbers in the paper and, conveniently, within the range of NTA) containing 1000 molecules that's only 10^{11} molecules, so it's hardly perturbing the bulk concentration. In this specific system, even when the clusters were in the 100nm size domain, it took hours for a crystal to appear, so there's plenty of time to look at how the clusters behave. After all, NTA is based on videos of the clusters moving around so it's possible for the human eye to interrogate the behaviour and learn new things. This allows us to start asking the question: "What's happening to the million clusters that do nothing while one of them decides to convert to a crystal form?"

It is important to note that experiments on these large clusters require extensive filtration to remove bits of junk that could cause other forms of cluster growth. A further important fact from the paper is that the 100 nm clusters could be repeatedly filtered through pores much too small to allow 100nm crystals to pass through, proving that they are not CNT-style nuclei. Quite how/why these clusters stay together is unclear. The theory of metastability (after all, these things should, thermodynamically, be forming large crystals) is challenging. But that's the point - if we can get lots of data on real clusters in real solvents, the theory will have a chance to catch up.

---

138 Shuyi Zong et al, *Molecular evolution pathways during nucleation of small organic molecules: solute-rich prenucleation species enable control over the nucleation process*, Phys.Chem.Chem.Phys., 2020, 22, 18663

I've mentioned a number of times that too many papers confidently assert that they have found *the* answer to crystallization, without acknowledging that crystallization is a property of the system and therefore spans a complex space. This paper makes no such claim:

> *"With respect to the important and still controversial question of the mechanism behind nucleation, the findings in this work could provide some new insights. However, since the nucleation process of crystals is really complicated and many factors could affect it, much more work needs to be done in the future to fully understand the nucleation phenomenon."*

Taking inspiration from the techniques in these bulk solution papers, plus the fancier liquid-TM techniques, and staying alert to the possibilities across crystallization space, maybe we can start to make more rapid progress in linking cluster science to everyday realities. A few more examples of solutes that do and do not (negative experiments have their value) provide clusters measurable via one or more of these techniques will start to give us a better idea of what's going on. Link these experiments to those designed to look at the cascade of clusters that can be created on, and swept off, crystals and we will surely make some rapid progress.

11.9.1 How stable is a cluster?

Small, loose clusters will tend to be statistically short-lived and large, tight clusters will be long-lived. That's not much of an answer but once again we have to recognize that there is a continuum of properties. A 50nm cluster with no molecular self-attraction would, if it could form by chance, disperse within µs via conventional diffusion. We know from the 50:50 polymorph experiments discussed earlier that some crystals can be heated to a temperature that allows full visible dissolution but must contain clusters which can survive for 10s of minutes and still preserve sufficient information about their preferred crystal form. The fact that NTA could be performed on OTBN clusters shows that they are around long enough to track and in any case we are told that the clusters can pass through nano-sized filters. Other papers cite "10s of seconds" for cluster lives. As we get closer and closer to clusters large enough to detect by turbidity, it seems that their lifetimes are essentially infinite.

Once the general question is asked, the specifics follow. Do "my" clusters (different solute/solvent combinations will be different) gently fade back to the background concentration, do they spontaneously break down into smaller ones, or do they grow each time two clusters collide? Do they love to be attracted to some surfaces, are they repelled by others? And when do they decide to tip over into being crystals?

To answer those questions in a coherent manner we will need to create cluster factories, discussed below.

11.9.2 Cluster theory problems

In KBI language, all we need to do is to understand the solute-solute interactions, G_{uu}, the solute-solvent interactions, G_{u2} and the solid-solute and solid-solvent interactions, G_{su}, G_{s2} plus any higher-order interactions such as "solid induced solute aggregation" just as with hydrotropes we had "solute induced hydrotrope aggregation". Historically, KB has been more comfortable with interactions in the liquid phase rather than at a solid surface. KB has also been happier staying away from phase changes where KBI become huge and, eventually, meaningless.

The complementary approach, density-functionals, is in a similar state - we know that it should be good but don't know how to apply it.

So, at the time of writing, we don't even have the beginnings of a coherent cluster language that can work across the many aspects that control multiple phenomena from simple crystallization to oiling out, via surface interactions with junk and/or crystals surfaces.

The common temptation is to carry on using the wrong approach. It is all too common to read things like "CNT is of limited value, but here's our analysis using it and, oh dear, we haven't learned very much". 100 years of using the wrong sort of language has not helped us very much. In contrast, every time I've seen KB theory applied to similarly confused areas it has brought great clarity and promoted insightful analyses, often based on data mis-analyzed by failed theories. There are enough hints from the modern KB literature that phase transitions can be tamed, and that liquid-surface interactions are tractable, to convince me that this is the intellectual way forward.

The good news is that computational techniques such as molecular dynamics work naturally with RDF and KB analyses and that powerful techniques such as SANS and SAXS map automatically onto KB. So I believe that although we don't yet have the theoretical tools to properly exploit the cluster approach there are enough hints from enough independent groups that it's a matter of time before they develop. Till then it's up to the experimentalists to take the lead in answering the question: "Clusters? So what?"

11.9.3 Clusters? So what?

This, for me, is the hardest question. What would we do differently in the lab if we believed in clusters rather than the other more limited theories?

The first step is to stop wasting resources on the classic approaches that have produced so little for so much effort. If you have good evidence that in your specific system CNT and, say, BCF are the correct models, that's great. Otherwise, just say no.

Next we need cluster factories - to make clusters when we want, where we want, at the sizes we want and with the degree of compactness we want.

Cluster factories via a direct route are hard, you can't create them by pipetting or jetting small drops. A picolitre sounds small, but a picolitre drop has a diameter of 12μm. A femtolitre isn't much better, it's still 1.2μm diameter. Attolitres pipettes are available and the sizes are getting there at 120nm, but we need zeptolitre drops to be doing interesting things at 12nm. It's exciting to see that there are exotic experiments of crystallization of zeptolitre metal drops (which happen to refute CNT) using TEMs to activate the pipette - but the conditions are hardly relevant to what we need.

The long tradition of looking at crystallization in confined spaces[139] has so far revealed rather little of use for those of us interested in macrocrystallization, for three interconnected reasons of size. The first is that the "microfluidics" is usually in the multi-μm scale, far too large to be directly related to the scale of clusters. Yet, second, there is a relatively large surface area of "wall" which provides a significant perturbation to the system - that's my problem with, say, the beautiful nanopore crystallization techniques. Third, when crystallization starts there is an immediate unsaturation in the remaining small volume, so subsequent steps cannot be investigated. However, given the variety and power of techniques used in this tradition it seems likely that there is potential for these approaches to develop cluster factories under exquisite control.

Those with access to cryo-TEM and LP-TEM can no doubt think up smart cluster factory experiments. What about those who prefer experiments in test tubes?

For those systems where liquid clusters of significant size (>10nm) are relatively stable and can survive nano-filters, then a brief period of secondary nucleation (sliding, tapping, stirring of smaller crystals) followed by nano-filtration would allow, first, cluster size analysis (DLS, NTA, SAXS, AUC...) and, second, rational experiments on crystal growth by injecting a known number of clusters of known sizes into a solution with, for example, a slightly higher saturation.

That paragraph contains a lot of assumptions. But they are explicit and refutable which means that testing them will provide some negatives for some systems and, I predict, plenty of positives for other systems. Crystallization is a property of the system so cluster factories will be a rich area for exploration only in some systems. My reading of the literature tells me that "some" means "a lot", others may disagree.

---

139 These techniques are, in general, answering very different and important questions often relevant to biomineralization and smart structures - of huge interest and importance, but not to this chapter. A wonderful review which shows the vast range of interesting science that is found in this domain (and makes explicit some of the limitations in terms of macroscale crystallization) is that of Fiona C. Meldrum and Cedrick O'Shaughnessy, *Crystallization in Confinement*, Adv. Mater. 2020, 32, 2001068.

If the experiments mostly refute what I currently believe then I will do what I always do, apologize in public for my errors and rewrite the relevant sections of this chapter.

If the experiments are encouraging then lots of exciting questions naturally follow. What controls the typical size of a cluster for a given solute in a given solvent? How long can they last in a quiescent test tube? How readily can they be filtered? Do they fragment on filtration? How readily do clusters aggregate and can aggregation be encouraged or inhibited on demand? What are the rules for attraction towards surfaces, and how much does that result in cluster-into-crystal growth or just creation of larger clusters that are swept off?

The answer to "so what?" is either: I apologize for being wrong or lots of fresh avenues for research open up.

11.10 In summary

Despite all the obstacles, crystallization is a routine task done on the small scale in synthesis labs and on vast scale in manufacturing. The problem is the sizeable pain and expense needed to get to a reliable process. There is no sign that the standard ways of thinking of crystallization offer any breakthrough technology that will transform the relative chaos into predictable order. High throughput, used as a substitute for thought and knowledge will not crack the problem. What we *can* do, and is standard in the best organizations, is apply solubility science to get quickly to a reasonably good process and, in parallel, to increase the odds of the right polymorph reliably crystallizing where we want and when we want.

The idea of a crystallization map seems to me to be a helpful way to think what might be going on in your specific system and, therefore, how you might rationally guide things in your desired direction. Such a map sees a continuum from classic crystallization of simple systems such as NaCl through an array of cluster-filled landscapes and near or actual phase separation into "oil" drops which may or may not themselves start to crystallize; along the way it accommodates production of seeds via a continuum of mechanisms, plus varying fates for those seeds within the bulk or on the surface of macrocrystals.

What is striking is that although individual aspects of this cluster picture are acknowledged to a greater or lesser extent, this has not resulted in a coherent language to explain the variety of effects. A crystal growing inside a loose cluster is not too puzzling, but a loose cluster *not* growing on the surface of a crystal *is* puzzling, at least in the context of commonly-held theories. My belief is that our traditional language is not adequate to unite this complex landscape while the language of fluctuations, Kirkwood-Buff and density-functionals theory will prove ultimately to provide the intellectual unity we need not just to understand but to predict and control.

Until we have a sound theory, the wisest choice seems to be to consistently question what types of clusters might be influencing which portions of our specific area of interest - and to use any technique (IR, Raman, NMR, AUC, DLS, NTA, SAXS...) which might provide insights into solute-solute, solute-solvent and solid-solute, solid-solvent interactions. When this is coupled to the best of solubility theory, coupled with high throughput characterisation and quantification then we get the most benefit for the least effort - not perfection, but a lot better than the alternative..

12 Sorption

Put any material into contact with a vapour and it will *ad*sorb onto the surface and *ab*sorb into the bulk. The more vapour, the more sorption. The shape of that curve from 0 to 1, i.e. the fraction of the saturated vapour pressure is called the sorption isotherm because each such dataset is produced at a single temperature. There are various conventions for naming that fraction, I will just use φ. For water sorption isotherms it is often shown as "a" or "a_w" (for water activity) or simply %RH for relative humidity (most of us don't know the difference). You can learn a lot from an isotherm about the interactions between the stuff doing the sorption, the *sorbent* and the stuff being sorbed, the *sorbate*[140]. As we shall see, you learn a lot more by creating a set of isotherms at different temperatures.

And, of course, you can learn lots of molecular details by comparing isotherms of different vapours - say water versus ethanol.

When we measure from 0→1 that is ad- (or ab-) sorption. Going from 1→0 is desorption. In simple systems, the isotherms are similar. But if the vapour itself strengthens the interactions at the surface, the desorption curve shows higher values for most of the isotherm. As is common with "up" and "down" curves, the difference can be called "hysteresis".

For most of this chapter we will focus on adsorption. Happily, the (new) theory behind absorption uses the same equations so even if your prime interest is absorption, keep reading.

12.1 It's a mess

Before 2021 this chapter could not be written, at least, not coherently. The reason is that sorption isotherms are a complete mess. To provide some sort of order, IUPAC have decreed that there are 6 types of isotherm. But after that it's the wild west. Anyone who cares to look carefully can easily find over 80 isotherm models/equations (I will use "model" to imply "equation" through this chapter). Some are explicit that they are merely fitting equations, and others use some plausible theory that links the isotherm to some core parameters.

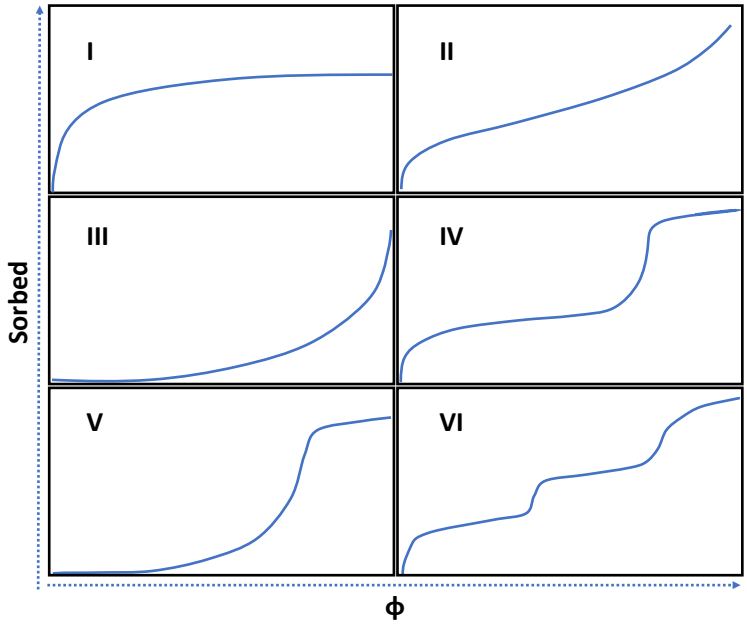
For those who want an expert, independent view of the situation, a review by Peleg[141] shows how dire the situation is. He has provided a rather powerful "merely fitting" model not so much in the hope that people will use it, but to emphasise that there is no correlation between "goodness of fit" and "link to

---

140 It is unfortunate that it is so easy to confuse the terms "sorbent" and "sorbate". But I can't change the terminology.

141 Micha Peleg, *Models of Sigmoid Equilibrium Moisture Sorption Isotherms With and Without the Monolayer Hypothesis*, Food Engineering Reviews (2020) 12:1–13. This paper is cited with full admiration for his consistent fight against sloppy thinking on isotherms and other issues (e.g. bad use of WLF) in food science.

the science behind the model". His model is "science free" (and, indeed, is technically flawed as $\varphi \rightarrow 0$ as it contradicts Henry's law), but provides excellent fits to a vast array of food isotherms where he is an expert.



When you have 80+ models to cover 6 types of isotherms that's a good indication that most or all of them are wrong or at least (and this includes the pure fitting models) have no firm connection to the underlying science. If any of them was right, then everyone would be using them. Readers who are familiar with BET and GAB will now say "Ah, everyone uses those so they must be correct". But as we shall see, they are just as wrong as the others.

It is quite a claim to say that after decades of hard work, with 6 isotherm types and 80+ models, there is no connection to the core science. But it is a true claim with an exciting upside. The upside is that any reader of this book who has paid some modest attention to Kirkwood-Buff will have no problem understanding the real, assumption-free theory which applies to all known isotherms because it's just plain old statistical thermodynamics which cannot be argued with.

As I have admitted in previous chapters, the downside of KB-style explanations is that it's sometimes hard to translate the pure theory into molecular interpretations. Happily, the new sorption theory allows us to create a set of models with fitting parameters that, in turn, link to the core molecular behaviours. But aren't these just a bunch more ad hoc models to add to the 80+ alternatives? No. They are derived directly from the pure theory and their limitations are explicit and known in advance.

Those who are familiar with BET and GAB, for example, will find that the model based on statistical thermodynamics is identical in form to those much-used models - which isn't surprising because they model a wide range of practical isotherms. What is different is that BET and GAB are based on assumptions that are completely unfounded (which is why BET/GAB papers are full of confusing discussions) while the fitting parameters from the new model have precise, assumption-free meanings.

Those who are familiar with cases of cooperative adsorption in porous particles, will know that BET/GAB-style models cannot work. This means that the core assumption-free data need to be fitted via a different set of parameters that capture the essence of cooperative adsorption.

Don't worry. It all turns out to be very simple. You will very quickly understand how to interpret *any* isotherm in broad KBI terms and then find an app that lets you use a fitting model based on stat-therm that provides core insights within the necessary assumptions and simplifications. Those wanting to dig deeper into the (surprisingly simple) theory can find the key papers by Shimizu and Matubayasi.[142]

12.2 The simple stat-therm

We are interested in the average number of molecules per gram sorbed onto our sample. Because this is stat-therm we use the angled brackets ‹› to show that we mean "ensemble average" and we use n for the number, so our plots are ‹n› versus φ.

12.2.1 Obtaining G_{s2}

It turns out that the standard isotherm plot is not quite showing pure stat-therm values. But if we plot ‹n›/φ versus φ we get a plot that shows, directly, G_{s2}, the KBI between surface, s, and sorbate, 2. That's amazing. Just a slight re-plotting and we get instant molecular insights. What is G_{s2}? It is a measure of how many sorbate molecules are on the surface *in excess of what you'd expect from random encounters*. If the sorbate is strongly adsorbed, then G_{s2} at that value of φ will be large, otherwise it will be small.

In a typical isotherm, G_{s2} will start off relatively large (the sorbate latches on quickly to the near-empty surface), decreases partly because the attraction decreases and partly because the expected average increases, so there's less special behaviour. At the hightest φ levels, G_{s2} might start to rise for reasons we shall shortly see.

If we plot φ/‹n› we have essentially the same information, it just emphasises different parts of the isotherm. However, if we take the derivative (we can do this straight from the raw data, we are *not* doing any fitting or making any assumptions) we obtain G_{22}/v. And that's it. From one plot and one derivative we have all the KBI information we need. For any isotherm. And, as it happens, it doesn't matter if we have adsorption or absorption, we get the same two KBI, directly.

But while G_{s2} sort of makes sense, G_{22}/v is trickier, more subtle and hugely important for understanding why the 80+ isotherms have led us astray.

12.2.2 What is G_{22}/v?

G_{22} itself isn't complicated, but it *is* suprising. It is simply the amount of sorbate around another sorbate more (or less) than you would expect from the random

142 A list of papers goes here

average. As we shall see, at low φ G_{22}/v is negative, meaning that you have fewer sorbate-sorbate interactions than you would expect from the average. But let's focus on why G_{22} is a big deal. The mysterious "v" will be explained later, for the moment we focus on sorbate-sorbate interactions which are hugely important yet, in common isotherm models, are *assumed not to exist!*

A typical model of an isotherm concerns itself, obviously, with how the sorbate interacts with the surface. For the much-used BET (Brunauer–Emmett–Teller) isotherm the story goes that the sorbate goes onto the surface which fills up to a monolayer, and once that monolayer is filled, further sorbates start to assemble into multilayers. For the theory to work, any sorbate-sorbate interactions in the monolayer plane are seen as insignificant and the binding of multilayers onto lower layers is seen as "weak".

If your interests are super-flat surfaces being probed by nitrogen at liquid nitrogen temperatures then it's possible to imagine that there's plenty of truth in this description. But as has often been pointed out, for real surfaces with molecules such as water, the assumptions are nonsense; one well-know quote from Rouquerol et al says "the BET model does not provide a realistic description of any known physisorption system". Despite this obvious fact, many generations of isotherms have been analysed using BET or it's slightly more complex version, GAB (Guggenheim- Anderson-de Boer) and people have solemnly reported numbers such as n_m the "monolayer adsorption number" which you can multiply by an area of each molecule to get a "BET surface area".

Countless papers compare and contrast things like BET surface areas and worry about modest differences of how the monolayer gets covered.

Sadly for BET and GAB, nature is statistical and molecules don't know that they should form neat arrays of monolayers. If this wasn't already obvious, G_{22} shows us that it should be. Remember, G_{22} is assumption-free stat-therm so we are not discussing a "model", this is how it is.

For a typical Type II isotherm analysed via BET or GAB, G_{22} starts negative and rises, in a straight line, to positive values at higher φ. The fact that it is a straight line is significant. The isotherm has a "knee" at the point which is attributed to the monolayer coverage - something that makes intuitive sense. But neither G_{s2} nor G_{22} shows any special "knee-like" behaviour. That's because the monolayer is a fiction.

G_{22} starts negative for the simple reason discussed in other KBI contexts - if one molecule is somewhere then another molecule can't be in the same place. It's the "excluded volume" effect, something at the same time obvious yet so often forgotten that "excluded volume effects" are sometimes seen as mysterious.

So, we start with few sorbate-sorbate interactions. Yet soon they build up till at a φ values that depends on the system, G_{22} becomes positive - sorbates love to associate with each other, entirely going against the assumptions of many isotherm models but being entirely obvious to any chemist.

If the sorbate is water, the temptation is to ascribe the interactions to the strong hydrogen bonding capabilities of water. Yet these positive G_{22} values occur for every isotherm I've ever analysed. The clue to this is the fact that both G_{s2} and G_{22} are derived from essentially the same curve, i.e. ‹n›/φ or φ/‹n›. It is the *surface* which is attracting the sorbate (so G_{s2} is positive) and it is the *surface* which is inducing sorbate-sorbate interactions. So although G_{22} only contains sorbates, it is the surface-induced sorbate-sorbate interactions that we are looking at.

So the very thing that BET and GAB reject - sorbate-sorbate interactions - is at the heart of every isotherm.

Now we understand the significance of G_{22} we can focus on the fact that what we measure from the re-plotted isotherm is G_{22}/v. We will quickly see that without v none of this would make sense. The definition of v is the "interfacial volume" per unit quantity of sorbent, which is the region in which the local concentration of sorbate deviates from the reference systems (the bulk sorbate gas exterior and the impenetrable solid sorbent interior). At very low φ values on a relatively smooth substrate this can be imagined as the volume of a monolayer of sorbate (though, of course, we never have such a monolayer!). At higher φ values v will be a few nm thick as there are statistical multilayers. On complex surfaces there is no unambiguous visualisation of v, but it is still a valid thermodynamic quantity.

For the moment, all we need to grasp is that because v is a volume per unit quantity of sorbate, it captures the obvious fact that some sorbates have a larger available surface area (per unit quantity) then others, and if, for simplicity, you multiply that surface area by a molecular thickness, you have the interfacial volume, v.

The value of v can be measured in principle at any value of φ via sophisticated analytical probes and can be well estimated by computer simulation. So it is a value "knowable in principle" but for most of us never knowable in practice. This is unfortunate, but it is just the way it is. However, there is one value of v, let's call it v^0, which we can understand and use to our advantage. It goes along with $G_{22}{}^0$ which is the value of G_{22} when $\varphi \rightarrow 0$. A final value is $G_{s2}{}^0$ with an obvious meaning. We will return to these important values later.

12.3 ABC

Our curves of G_{s2} and G_{22}/v give us direct insights, across the whole φ range as to what is going on. Because they are assumption-free we can work with them in any isotherm situation. But humans aren't so good at comparing/contrasting whole curves. We like to have numbers so that *this* sorbent has *this* key parameter value with, say, water while *that* sorbent has *that* key value.

So now we will introduce some simplifying assumptions so we can derive human-scale numbers. In each case, the assumptions are reasonable statistical thermodynamics so they don't make any a-priori assumption about mechanisms. We don't impose mechanisms, we simply use valid statistical approximations. If the data fit nicely then the approximations make sense. If they don't then there are other, equally reasonable sets of approximations that can be made. We will come to those later. First we will look at typical day-to-day isotherms such as we find in water sorption on foodstuffs - one of the huge areas of isotherm studies.

It is no coincidence, as we shall see, that these use BET and GAB to fit them.

It is common in thermodynamics (and entirely respectable, think "virial coefficients") to say that a key parameter can be expressed as a polynomial series. It turns out that G_{22}/v can be nicely approximated by (stopping at C, terms in D and φ^2 etc. could be included for more difficult isotherms):

$$\frac{G_{22}}{v} = B + \frac{C}{2}\varphi$$

Equ. 12-1

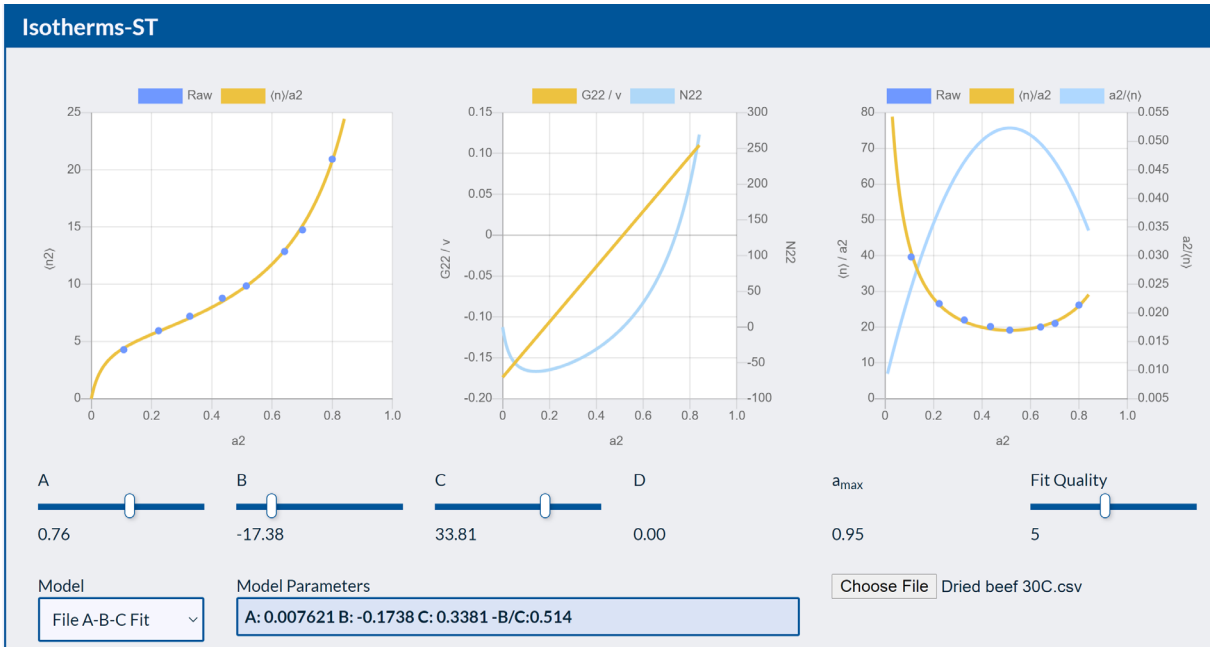
The B term describes sorbate-sorbate (doublet) interactions (that's what G_{22} is!) and the C term with its dependency on φ describes triplet interactions. We can imagine a D term to describe quadruplets. For many day-to-day isotherms we don't need the D term, so the discussions focus on ABC.

From this equation and a bit of simple integration, we can obtain the fitting formula where the A term describes pure sorbate-surface interactions, i.e. is related to G_{s2}:

$$\langle n \rangle = \frac{\varphi}{A - B\varphi - \frac{C}{2}\varphi^2}$$

Equ. 12-2

The app does all the basic work for you if you give it an isotherm dataset in a simple .csv format:



App 12-1 https://www.stevenabbott.co.uk/practical-solubility/Isotherms-ST.php

The first graph shows the isotherm itself (instead of φ the app uses a2 for the actiity of water), the middle graph shows G_{22}/v and the related value of N_{22} which is the excess number of waters compared to the average. The third graph shows both graphs, ‹n›/φ and φ/‹n› with the first being, essentially G_{s2}.

The key output parameters are also provided. If we ignore a few issues of units and standard states (we will address those shortly) we can say what each of the parameters means.

1/A is G_{s2}^{0}, i.e. a measure of the strength of the s↔2 interactions when there is just a single molecule involved ("infinite dilution"). It is the equivalent to using an AFM probe with the sorbate on its tip, looking at the attraction of a single molecule on an otherwise empty surface.

B is G_{22}^{0}/v, and is typically negative because of excluded volumes. The simplest interpretation, which works well for typical sorbates on typical surfaces, is that it is *entirely* the excluded volume, known from the sorbate's molar volume, divided by v. The significance of this will become clear soon.

C is the slope of the G_{22}/v curve. The larger its value, the stronger the surface-induced sorbate-sorbate interactions are.

-B/C is a useful guide because it is the value of φ when G_{22}/v becomes positive. The lower this value, the stronger the interactions.

12.3.1 From ABC to BET & GAB

Earlier I claimed that the assumptions behind the BET and GAB models were incorrect. Now we can find out, via pure stat-therm, why this is the case.

The BET model uses n_m the assumed monolayer equivalent coverage and C_B the BET constant about which most users of the theory are a bit vague. GAB takes BET and adds an extra factor K, which is 1 for BET. The GAB formula (or BET if you set K=1) is:

$$\langle n \rangle = \frac{n_m C_B K \varphi}{(1 - K\varphi)\left[1 + (C_B - 1) K\varphi\right]}$$

Equ. 12-3

Although it is not obvious, the functional form is identical to the ABC formula, so BET/GAB are essentially the same as ABC. The difference is that each value of A, B & C has a specific, well-defined, assumption-free meaning while n_m, C_B and K are based on a mechanism (monolayers etc.) that happens to be wrong.

Happily we can re-use the entire valuable corpus of BET/GAB data by translating their parameters into ABC. It's just a bit of slightly complex algebra:

$$A = \frac{1}{n_m C_B K} \quad B = \frac{2 - C_B}{n_m C_B} \quad C = \frac{2K(C_B - 1)}{n_m C_B}$$

Equ. 12-4

For respectable BET, C_B has to be "large", at least greater than 80, so for anyone who has a good BET plot the C_B terms in B more-or-less cancel out so $B = 1/n_m$. It therefore turns out that the "BET monolayer coverage" is actually capturing v^0/G_{22}^0, the interfacial volume divided by the sorbate-sorbate KBI, both at $\varphi \rightarrow 0$.

And we can see why the "BET constant" has never been used with conviction. On its own it has no special meaning, but combined with n_m it equals 1/A, which means that it is related to G_{s2}, i.e. the measure of how strongly an isolated sorbate is attracted to the surface. For GAB the K factor is also part of the link to 1/A.

A quick glance at the conversion equations shows that for BET when K=1 and C_b is large, C is identical to A, just $1/n_m C_B$. This isn't surprising - BET is a two-parameter model so A and C are identical. C becomes more significant for GAB because now K can now be interpreted as a descriptor of the strength of sorbate-sorbate-sorbate interactions - a large K means that the sorbate clusters faster than a low K.

It is easy to look at the equations and ask "Well, if GAB and ABC are equivalent, what's the fuss?" The fuss is that there have been decades of confused discussion around n_m, C_B and K. We now see that confusion was guaranteed, because the constants did not mean what the model assumed they meant, so interpretations based on the wrong model could not be coherent.

The biggest confusion of all arises over "BET surface area".

12.3.2 BET surface area

If n_m is a monolayer coverage and if each molecule takes up an area of σ, then you can instantly calculate the total surface area of your sample as $n_m\sigma$.

If your isotherm is with nitrogen at 77°K and if your surface isn't too complex, then such a measure of "the" surface area seems to make intuitive sense. And if you can find alternative measures of the surface area, there's often a good match.

If your isotherm is with water at 30°C and your surface is dried beef, it seems rather unlikely that the calculated area is the "real" value. Indeed, if you put the same dried beef into a nitrogen BET device, the area may differ by an order of magnitude.

And here's the big problem with BET surface area. Entire communities take a theory that is not so bad at 77°K with nicely boring nitrogen molecules and try to derive meaning from experiments on complicated surfaces with complicated molecules such as water. When they obtain differing values they have to create ad hoc explanations of what's going on. When you start with a model that happens to be wrong, but which works adequately in its original domain and apply it to systems way outside the original domain, the results can only be confusion.

So let's un-confuse the situation. We now know that n_m is, in fact, v^0/G_{22}^0 and that there isn't, in most real-world cases, a monolayer coverage happening. We also know that $n_m\sigma$ is often capturing something significant about two surfaces - a sample of particles with a diameter around 50nm will show a higher value, maybe 10x higher, than a sample of 500nm particles. So can we just calculate σ v^0/G_{22}^0 and get our own surface area?

Yes and no. Let's take a relatively simple surface where we can plausibly define v^0 as being the real surface area times the thickness of a monolayer of the sorbate. And let's take the default standard KBI assumption that G_{22}^0 is the excluded volume, i.e. more-or-less the molar volume of the molecule. Instantly, v^0/G_{22}^0 becomes the number of molecules that could fit into a virtual monolayer made up of sorbates that have no interest in any interaction with each other. Multiply that by σ and we have our surface area. Our 50nm particles will have

a larger v^0 than our 500nm particles, so we are capturing something that represents a surface area.

But here's the difference from "BET surface area". We are insisting that the chain of logic from v^0/G_{22}^0 to a surface area is a plausible one under simple conditions. If the isotherm is with nitrogen at 77°K this might well be a real surface area that's being measured. As soon as we have something complex like a food surface, it isn't at all clear that v^0 is meaningfully one molecule thick, nor is it clear the G_{22}^0 should be the excluded volume. There are plenty of food isotherms where B turns out to be positive, i.e. the sorbent-sorbate-sorbate effects are so strong that even at infinite dilution the excluded volume effects are overcome.

The fact that we cannot know, just from a simple isotherm, what v^0/G_{22}^0 means in terms of molecules and areas is liberating, not restrictive. Far too much effort has gone into "explaining" why two surfaces have different BET surface areas, when, in fact, there are multiple possible explanations based either on v^0, G_{22}^0 or both. If we have independent evidence around either of those factors, then we can use it. If, as is usually the case, we just have B, then we acknowledge that we cannot say profound things about surface areas. Given that most of us would be hard pressed to define "the" surface area of a complex food or a porous carbon, it is better to not pretend that we can.

12.3.3 ABCD...

If our ABC fit is poor, we can choose to add a D parameter and interpret the fitted results through strengths of quadruplet sorbate-sorbate interactions. In my experience this is just about plausible. We could add E, F, G... parameters to fit ever more complex curves. While this remains mathematically acceptable, because it's just higher term virial coefficients, we all know that multiple fitting parameters provide rapidly diminishing returns in terms of what we crave most - understanding of our system.

In any case, the ABC... approximation assumes that nothing especially interesting or significant is happening within the system. For systems where we absolutely know that some really interesting stuff is happening, let's use an approximation that explicitly takes it into account. Welcome to cooperative isotherms.

12.4 Cooperativity

The Type V isotherm could be described via some complex polynomial, but the fitting parameters would provide no insights or allow sensible comparisons between different isotherms. Just as the ABC fitting makes use of the minimum possible assumptions, we can fit a cooperative isotherm assuming that there is some A term to capture G_{s2}^0, a B term to describe the general isotherm

when behaviour is relatively simple then an M term to bundle together all the cooperative interactions involving a cluster of m+1 molecules. We already have that if we consider C captures 2+1, i.e. triplet interactions, but here we are interested in serious clustering where m might be, say, 10.

After some stat-therm wizardry we find that there is a gratifyingly simple formula:

$$\langle n \rangle = \frac{B\varphi + mM\varphi^m}{A + B\varphi + M\varphi^m}$$

Equ. 12-5

Remember that this is a (thermodynamically plausible) fitting function, while the essentials, the G_{s2} and G_{22}/v, remain universal. We are free to interpret the curves directly, or we can choose to mine the parameters of the fitting function to be able to compare and contrast different isotherms.

As we already know about A and B, we just need to discuss M and m.

M is a measure of how much cooperativity there is - the larger M, the more cooperative the behaviour compared to the baseline via B.

The m parameter tells us how many molecules exist in a typical cluster. As this is statistical thermodynamics, we know that clusters will extend from 1 to many with some sort of probability distribution, and the simple fitting function cannot tell us about the shape of that distribution. Instead it gives us a reasonable idea that *this* sorbate on *this* material has a typical cluster of 8 molecules and *that* sorbate on *that* material has a typical cluster of 12. How you then interpret those values is up to you.

12.4.1 Types IV and VI

In each of these cases there are multiple sub-clusters so the fitting function needs sums of terms similar to the Type V equation. The main outcome is sets of M and m values which describe the relative significance of that specific clustering and the number of molecules in the cluster. If any reader wants to get serious with such isotherms, the relevant apps are available.

12.4.2 Type III

At the time of writing, this "condensation" type of isotherm has no convenient fitting function to provide meaning. The reader is free to interpret the G_{s2} and G_{22}/v values as they wish.

12.5 The other end of the isotherm

Using Inverse Gas Chromatography (IGC) it is possible to get high-quality desorption isotherms from various probe molecules on various "interesting" particulate materials. It is standard practice to analyse these isotherms using a special type of BET analysis. The details are described in my IGC eBook, https://www.stevenabbott.co.uk/practical-chromatography/the-book.php. At the heart of the analysis is the fact that some materials have "high energy sites" where the probe molecules bind especially tightly. These are often a tiny fraction of the general surface but can have significant effects on how the material behaves when used in formulations.

The IGC analysis creates and AEDF, Adsorption Energy Distribution Function, derived by assuming that the desorption curve is a function of different BET monolayer capacities and BET constants.

Because we know that the BET analysis happens to be incorrect, the obvious question is whether the ABC fit can analyse these desorption curves. The answer is "yes, but". Most of the curve is well-described, and the ABC parameters can be quantified and analysed in the usual way.

But frequently the isotherm shows a low φ slope that cannot be properly fitted with ABC. To deal with this we need another plausible model, and the simplest, meaningful model is to posit one or more high energy sites each of which follows the simplest possible isotherm, the Langmuir, which says there is a surface which gets filled with molecules, end of story.